Statistical Policy Working Paper 10

Approaches to Developing Questionnaires

Statistical Policy Office
Office of Information and Regulatory Affairs
Office of Management and Budget

November 1983

MEMBERS OF THE FEDERAL COMMITTEE ON STATISTICAL METHODOLOGY

(November 1983)

Maria Elena Gonzalez (Chair)
Office of Information and
Regulatory Affairs (OMB)

Barbara A. Bailar
Bureau of the Census
(Commerce)

Norman D. Beller
National Center for Education
Statistics (Education)

Yvonne M. Bishop
Energy Information
Administration (Energy)

Edwin J. Coleman
Bureau of Economic Analysis
(Commerce)

John E. Cremeans
Bureau of Industrial Economics
(Commerce)

Zahava D. Doering
Defense Manpower Data Center
(Defense)

Maria D. Eldridge
National Center for Education
Statistics (Education)

Daniel H. Garnick
Bureau of Economic Analysis
(Commerce)

Charles D. Jones
Bureau of the Census
(Commerce)

Daniel Kasprzyk
Bureau of the Census
(Commerce)

William E. Kibler
Statistical Reporting Service
(Agriculture)

Thomas Plewes
Bureau of Labor Statistics
(Labor)

Fritz J. Scheuren
Internal Revenue Service
(Treasury)

Monroe G. Sirken
National Center for Health
Statistics (Health and
Human Services)

Thomas G. Staples
Social Security Administration
(Health and Human Services)

Statistical Policy
Working Paper 10

Approaches to Developing Questionnaires

Prepared by
Subcommittee on Questionnaire Design
Federal Committee on Statistical Methodology

Edited by
Theresa J. DeMaio
Bureau of the Census

MEMBERS OF THE SUBCOMMITTEE ON QUESTIONNAIRE DESIGN

Dawn D. Nelson (Chair)
Bureau of the Census (Commerce)

Deborah H. Bercini
National Center for Health Statistics (HHS)

Theresa J. DeMaio
Bureau of the Census (Commerce)

Richard W. Dodge
Bureau of the Census (Commerce)

Gemma M. Furno
Bureau of the Census (Commerce)

Maria E. Gonzalez* (ex officio)
Office of Information and Regulatory Affairs (OMB)

Janice Olson
Social Security Administration (HHS)

Anitra Rustemeyer Streett
Energy Information Administration (Energy)

Ronny Schaul
Bureau of Labor Statistics (Labor)

Margaret Weidenhamer
Statistical Reporting Service (Agriculture)

Additional Contributors to the Report on Approaches to Developing Questionnaires

Catherine J. Baca
Bureau of the Census (Commerce)

Carol M. Utter
Bureau of Labor Statistics (Labor)

*Member, Federal Committee on Statistical Methodology

OFFICE OF INFORMATION AND REGULATORY AFFAIRS

Christopher DeMuth, Administrator

Thomas D. Hopkins, Deputy Administrator for
Regulatory and Statistical Analysis

Dorothy M. Tella, Chief Statistician

Maria E. Gonzalez, Chairperson
Federal Committee on Statistical Methodology

PREFACE

The Subcommittee on Questionnaire Design was formed by the Federal Committee on Statistical Methodology to address the general topic of questionnaire design. The Subcommittee focused on a review of methods used in developing questionnaires. The working paper discusses approaches to devising questionnaires in three broad areas: tools for developing questions, procedures for testing the questionnaire draft, and techniques for evaluating the questionnaire.

While the report is intended primarily to be useful to Federal agencies that develop questionnaires, a broader audience may also find the report of interest. Seminars and meetings will be organized to discuss the topics addressed by this subcommittee with Federal agency personnel.

The Subcommittee was chaired by Dawn D. Nelson, Bureau of the Census, Department of Commerce. As a subcommittee report, this document does not necessarily represent the views of the Office of Management and Budget.

ACKNOWLEDGMENTS

This report represents the collective effort of the Subcommittee on Questionnaire Design. Although all members of the subcommittee reviewed and commented on the entire report, individual members were responsible for preparing various chapters. Chapters 4 and 11, however, were prepared by persons who were not members of the subcommittee. The names of the authors of the respective chapters appear below.

Chapter	Author
1	Theresa J. DeMaio
2	Anitra Rustemeyer Streett
3	Margaret Weidenhamer
4	Catherine J. Baca
5	Dawn D. Nelson
6, Section I	Gemma M. Furno
6, Section II	Janice Olson
7	Anitra Rustemeyer Streett
8	Deborah H. Bercini
9	Theresa J. DeMaio
10	Richard W. Dodge
11	Carol M. Utter

The following persons also deserve special recognition for their role in assisting the work of the subcommittee. Our work was initially guided by Naomi Rothwell who served as the chair from November 1980 until her retirement from government service in March 1981. Maria Gonzalez worked with the subcommittee throughout the development of the report and provided a link with the Federal Committee on Statistical Methodology (FCSM). Various members of the FCSM provided advice on our work at different stages, and Barbara Bailar and Zahava Doering supplied comments on the complete report draft. Much of the report material was also reviewed by Thomas Jabine who provided encouragement for our work.

We especially appreciate Theresa DeMaio's contribution in sharpening the focus of the report by effectively organizing and editing the material. Also, we are grateful to Laura Taylor, Vicki Horton, Debbie Barnett, and Cathleen Tyson of the Bureau of the Census for typing and assembling the many drafts of this report. The Bureau of the Census also provided the funding for publication preparation and printing.

CONTENTS

Part I

Chapter 1: Overview 3

Part II: Tools for Developing Questions 11

Chapter 2: Unstructured Individual Interviewing 13
Chapter 3: Qualitative Group Interviews 21
Chapter 4: Participant Observation 29

Part III: Procedures for Testing the Questionnaire Draft 41

Chapter 5: Informal Testing 45
Chapter 6: Formal Testing 57

Section I. Pilot Studies 57
Section II. Split Sample Testing 70

Part IV: Techniques for Evaluating the Questionnaire Draft 89

Chapter 7: Investigating Respondent's Interpretations of Survey Questions 93
Chapter 8: Observation and Monitoring of Interviews 101
Chapter 9: Learning From Interviewers 119

Section I. Interviewer Debriefing 119
Section II. Structured Post-Interview Evaluation 124

Chapter 10: Using Record Checks 137
Chapter 11: Response Analysis Surveys 151

References 159

Part I

Chapter 1
Overview

I. INTRODUCTION

Formulating a series of questions to obtain the answers to a set of data needs may appear to be a relatively simple task; however, constructing a questionnaire that will elicit accurate information from most respondents interviewed is more complicated than it may seem. For example, a seemingly simple question concerning vehicle ownership--How many cars do you own?--may appear to convey all the information necessary for respondents to answer it and to mean the same thing to respondents, survey designers, and data users alike.[1] However, upon reflection, such a question is not as clear as it seems. The word "car" may or may not be intended to include such vehicles as vans, campers, motorcycles, tractors, and snowmobiles; "you" may or may not refer to household or family members as well; "own" may or may not include vehicles which are leased or are in the process of being bought.

Questionnaire designers need to consider many factors during the process of creating a questionnaire. For example, will every question be interpreted in the same way by most respondents? If not, the data might not provide the information required by the questionnaire designer. Or, for another example, can respondents remember whether or not events of interest to the questionnaire designer have occurred within a given time frame, and if so, can they recall the details of those events accurately?

Some generally accepted rules exist for wording, sequencing, and formatting questionnaires and can be used to guide a questionnaire designer in constructing an initial draft of a questionnaire. Yet the development of any particular questionnaire is unique. Refinement is necessary to ensure that any questionnaire used in the field will produce sufficiently accurate results. In the example described above, for instance, testing of the question would reveal the ambiguities inherent in it and lead to the development of a question more likely to meet the data requirements.

The purpose of this report is to present a series of tools and tests which are useful in the initial drafting and subsequent refinement of a survey questionnaire, to explain their applicability to questionnaire design, and to describe the mechanics of implementing them. Numerous examples of these techniques are also provided to illustrate the points made. Although the

[1]This example was adapted from one described by Biderman et al. (1982).

Principal Contributor: Theresa J. DeMaio

focus is on survey questionnaires, many of the techniques are applicable to the development of data collection forms for administrative and other purposes.

Many of these techniques are relatively simple, inexpensive ways to improve the quality of a questionnaire. For the most part they are appropriate for developing survey questionnaires regardless of the type of information being collected (e.g., factual, behavioral, opinion, or knowledge), the method used to obtain it (e.g., mail, telephone, personal visit, or a combination), or the type of reporting unit (e.g., households, individuals, farms, or establishments). Used appropriately, these techniques should result in more efficient use of resources, reduced respondent burden and nonsampling error, and better realization of a survey's objectives. Maximum effort is justified during the developmental stage, because once a questionnaire is in use, problems are costly or impossible to correct. The time and money spent in developing a questionnaire should be repaid by collection of more relevant, better quality data.

II. AUDIENCE FOR THE REPORT

This report was written primarily for questionnaire designers in Federal agencies. While this does not limit the report's use by others, it may explain the focus and choice of materials for illustration. It is hoped that those who have relatively little experience in this area will benefit from exposure to the techniques available for questionnaire development and how to use them. Even more experienced questionnaire designers may not be familiar with all the techniques and they may find the report useful as a reference. The report may also be helpful to persons who do not design questionnaires themselves, but who work in agencies that sponsor surveys to be conducted by private contractors or other government agencies. It is hoped that circulation of this report will promote increased familiarity with some of the less frequently used approaches and encourage use of all the techniques, thereby improving the relevance and quality of the data collected by the Federal Government.

III. ORGANIZATION OF THE REPORT

The approaches described in this report are divided into three sections: tools used to develop questionnaires, tests conducted to examine questionnaires, and techniques for evaluating questionnaires during the testing and developmental work. The order of presentation does not imply that the tools and tests must be used in a step-by-step order to develop a good questionnaire. It would be too costly, time-consuming, and inefficient to use every technique in the development of a single questionnaire. Moreover, each technique has strengths and weaknesses (in terms of cost, time, and resource requirements, and questionnaire design issues for which it is relevant) that render it appropriate or inappropriate for a given purpose.

Within each chapter, an attempt is made to clarify when and how the technique can be used most appropriately. The topics discussed in each chapter follow the same general outline: I. Introduction; II. Method--A. Personnel and Skill Requirements; B. Selection of Respondents, C. Preparation;

D. Operation; E. Time Considerations; F. Cost Considerations; G. Mode of Data Collection; and III. Examples.

Those who use this report should also be aware of the data collection requirements imposed on agencies by the Paperwork Reduction Act of 1980 or any Federal regulations which supersede this Act. A discussion of the current requirements monitored by OMB is contained in Statistical Policy Working Paper 9, "Contracting for Surveys" (Office of Management and Budget, 1983).

IV. BACKGROUND: OVERVIEW OF THE QUESTIONNAIRE DESIGN ISSUES

The focus of this report is the development and evaluation of questionnaires rather than the drafting and design of the questionnaire itself. However, in order to provide a framework for understanding the relationship between the development and evaluation process and questionnaire design issues, a brief description of the general issues is presented here. This is intended to provide the reader with some understanding of why the techniques that are the subject of this report are important components of the questionnaire development process. (However, since it is only an overview, readers unfamiliar with the topic should refer to other sources for a more detailed treatment of of the issues--e.g., Payne, 1951; Sudman and Bradburn, 1974, 1982; Dillman, 1978; Bradburn and Sudman, 1979; Schuman and Presser, 1981; Turner and Martin, 1984.) In the chapters that follow, connections will be made between the techniques being described and the questionnaire design issues they are suited to address.

The primary questionnaire design issues addressed in this report are content, question wording, question sequencing and flow, and questionnaire administration. Each of these is described briefly below; several other issues of secondary importance are described following this section.

A. Content

Decisions concerning what to include and exclude from a questionnaire and still meet the survey objectives are crucial. The analysts and data users should be consulted as early as possible in the process of specifying the subject matter. If an aspect of the problem is overlooked entirely, questions which would allow a fuller understanding of the subject of the inquiry may be omitted. For example, a questionnaire about child care arrangements could provide inaccurate information if the designer assumed that all parents make explicit and formal arrangements for such care when informal arrangements also exist.

Alternatively, the content of a questionnaire can be limited by the type of data that can be collected--respondents may not be sufficiently knowledgeable to provide accurate responses to all questions. For example, in a survey of housing quality, a measure of floor space may be required; however, respondents may not be able to provide that information accurately. Some respondents may admit that they cannot answer the question, but others will provide inaccurate responses rather than acknowledge their ignorance. The extent to which people are able to answer the questions presented to them affects the quality of the data that are collected.

One type of request frequently made of survey respondents, about which knowledge and accurate recall are particularly problematic, concerns the recollection of whether specific types of events occurred and, if so, when they occurred. Survey researchers (or, for that matter, cognitive psychologists) have little information about the process, limits, etc., of human memory and how people place events in time when they are asked to recall the occurrence of a particular event. Asking respondents about events which occurred during reference periods of different lengths, and including examples which provide memory triggers as part of the question, have been used to increase the accuracy of recall data. To elicit the most accurate information possible, careful attention should be given to formulating such questions.

In addition to not being able to answer questions they are asked, respondents, for various reasons, may not want to answer some of the questions included in a survey. They may feel that the information being solicited is sensitive--that some harm will come to them if they report some fact (e.g., use of illegal drugs), that they will be embarrassed by divulging certain information to an interviewer (e.g., inability to read), or that certain information is private and should not be disclosed to strangers (e.g., income). Effects on responses due to sensitive subject matter may be minimized through the sequencing of the questionnaire (see section on question sequencing and flow); however, a questionnaire designer must realize that the subject may be sensitive and take precautions to minimize response error or item nonresponse.

B. Question Wording

To provide comparable data from every unit in the sample, the survey questions must, as nearly as possible, present the same stimulus to all respondents. Several questionnaire design issues relate to this requirement.

Generally speaking, the vocabulary used in each question should be familiar to respondents and mean the same thing to most respondents. Regional variations in the meaning of certain words would make them inappropriate for use in a national survey--for example, "soda," "soda pop," "pop," and "soda water" all refer to carbonated beverages, but any one of them would be interpreted differently in different parts of the country. A respondent who does not know what a word means to the researchers will not be able to provide an accurate response to a question that contains it. Respondents may provide answers to questions, but those answers may not reflect the reality intended by the questionnaire designer.

A similar situation occurs with regard to the meaning that questionnaire designers and respondents attach to particular concepts used in survey questions. If a respondent, or some limited subgroup of respondents, does not interpret a question in the same way as it was intended by the questionnaire designer, then the answer will not be a valid measure of the survey designer's construct. Over the course of administering the question to the entire sample of respondents, ambiguity is introduced into the results, leaving the investigator uncertain as to what those results really mean. For example, take a situation in which respondents are asked to rate the seriousness of crime in their neighborhood. Even respondents living next

door to each other may have different concepts of the boundaries of their neighborhood, which might affect their ratings. If the neighborhood were defined for them, differences in the ratings would reflect factors unrelated to conceptual differences in the geographic area covered. Investigators who are unaware of differences between their own "frame of reference" and the variety of "frames of reference" existing among respondents may interpret the results in ways that do not reflect reality.

Another question wording issue involves the response categories that are presented to respondents in fixed alternative or multiple choice questions--how many options should be offered, how they should be ordered, whether they should be presented in a forced-choice or open-ended format, whether a "don't know" option is presented. Decisions made in this area may affect the quality of the data obtained in the survey, since the answers provided by respondents will be distributed differently, depending on the alternative response categories offered.

The length of the questions in the interview is another issue related to question wording. An extended introduction to a question may afford the respondent time to think about the issues involved before giving an answer, thereby potentially providing more carefully considered and more accurate results. On the other hand, longer questions add to the length of the interview and may contribute to respondent fatigue, inattention, or confusion. Different types of respondents may react in different ways to long questions, introducing a systematic bias into the results.

C. Question Sequencing and Flow

Another set of issues in questionnaire construction concerns the order in which the questions are presented to respondents. Even if questions are worded so they mean the same thing to all respondents, response biases or problems in administration of the questionnaire may result from the way the questions are sequenced.

One such problem involves the context imposed by the previous question or perhaps a set of questions contained earlier in the questionnaire. Such questions may invoke a particular mind set in the respondent's consciousness which may not reflect the way he or she thinks about a certain topic in other settings. A respondent may thus answer survey questions differently, depending on the order in which they are presented. For example, a person may respond one way when asked to evaluate the overall quality of the neighborhood if a question has just been asked about the street lights in the area. The rating of the neighborhood may be influenced by an opinion of the street lights, even if street lights are not an important criterion for determining neighborhood quality. If the two questions were reversed, however, the rating of neighborhood quality might be different.

Another consideration is the location of so-called sensitive questions--questions considered intrusive or damaging to respondent self-esteem. Placing these questions late in the interview so they are asked after some degree of respondent confidence has been established may minimize refusals and response problems introduced by the nature of the subject matter. On the other hand, leaving such questions until the end may risk superficial

answers due to respondent fatigue. At the very least, such questions should be located where they fit logically in the flow of the questionnaire and, if necessary, be approached gradually through related, but less threatening, questions.

The overall flow of the questionnaire deserves attention in the questionnaire design process, since it too may affect the quality of the data that are collected. If too many items are included in a list, the amount of thought given to each response may decline towards the end because of respondent fatigue. Excessive consecutive questions with the same type of format may condition the respondent to "acquiesce" unthinkingly with the same answer to each question (e.g., to yes/no type questions). If questions about the same topic are included in several different places in the questionnaire, a respondent may become confused by perceived redundancy or hostile because of perceived carelessness and treat the survey interview with less seriousness than the investigator would like. Thus, for many reasons, the flow of the questionnaire is an important element of questionnaire design.

D. Ease of Questionnaire Administration

In designing a questionnaire, the ease with which the questionnaire can be used by the interviewer/respondent is an important consideration. One aspect of questionnaire construction involves the placement of instructions. The extent to which interviewers/respondents are required to flip through the questionnaire, refer to previous answers that are not readily accessible, etc., should be minimized. The harder it is for the interviewers to determine the flow of the interview, the more chances for introducing interviewer error, item nonresponse, and respondent frustration.

E. Other Design Issues

Several other elements in the design of the survey may be relevant to the construction of the questionnaire. These issues, which are related to procedural decisions and format of questionnaires, are described in this section; however, they are of secondary importance in this report.

External constraints imposed by cost, time, or OMB respondent burden requirements may dictate the length of the survey interview, thus limiting the amount of information that can be obtained in the questionnaire. This may affect the number of questions that can be included and, therefore, the number of topics included in the questionnaire or the amount of detail obtained about particular topics. To some extent, the types of questions that are included can also be affected. For example, time-consuming techniques such as randomized response or card sorting might have to be eliminated if they add too much time to the length of the interview.

The criteria for selecting survey respondents should also be considered in designing a questionnaire. A survey involving responses from every eligible household member may require questions to be worded differently than for a survey in which a single respondent answers questions about each household member. In addition, the sequence of questions may have to be altered slightly to accommodate different types of respondents.

The method of data collection may also influence the design of the questionnaire. Questions employing visual aids, which are helpful in face-to-face interviews, are obviously not feasible for use in telephone interviews. The differences between modes of interviewing in the dynamics of interaction between interviewer and respondent may also suggest alterations in the types of questions that are used to obtain data of comparable quality in different interviewing modes. For example, mail questionnaires might be more successful in obtaining sensitive information than either mode involving direct interaction with an interviewer, who might be perceived as judgmental of respondents' answers.

The unit of analysis for the data also affects the structure of the questionnaire or the type of information that is collected. If the objective of the survey is to compile data on families, information need not be collected about unrelated household members. If, however, data for households are required, information about unrelated household members would be needed. The provision of data for individuals, for specific population subgroups such as food stamp users, or for a combination of different units, may require alterations in the order or wording of questions.

For surveys in which each unit is interviewed more than once, the number of interviews in the sequence for each sample unit and the length of time between interviews may influence some aspects of the questionnaire. For example, the amount of elapsed time between contacts affects the length of the reference period used in asking respondents to recall events. If data are collected in a series of interviews within a specified time period, the number of interviews conducted within that time period may affect the length of each interview. That is, the same amount of information that is collected quarterly could be obtained in only three longer interviews per year.

Format is also an important issue in the design of the questionnaire. Concerns in this area relate to the appearance of the questionnaire--color or kind of paper, size or style of type, method of data processing, method of questionnaire administration, etc. These issues may affect the quality of the data by influencing how well respondents or interviewers are able to follow the instructions and answer the questions. Concerns about the relationship between appearance and answering a questionnaire correctly are, however, in a different realm than the previously described issues which relate to the meaning of the questions and respondents' ability to answer them accurately. The topic of format is not addressed directly in this report, although some of the techniques described here can be useful in this regard.

V. SUMMARY

The words "may" and "might" have deliberately been used throughout this description of questionnaire design issues. Many of the issues that have been raised here may, but do not necessarily, cause problems for questionnaire designers. Although progress has been made in the last several years in identifying sources of nonsampling error and in measuring its extent, guidelines for eliminating its existence through systematic rules for questionnaire design have not been forthcoming. Efforts to construct guidelines involve evidence based on individual cases and the extent that these

guidelines can be applied to questionnaires involving different subject matters, respondent populations, or survey designs (e.g., one-time vs. repetitive surveys) is not clear. Some issues are more clear-cut than others. For example, in the area of question wording, it is generally accepted that questions which "lead" the respondent in one direction or another should be avoided. Even in this instance, though, the determination of whether a particular question "leads" the respondent may be a subjective one. Moreover, a questionnaire designer may deliberately use "leading" questions to meet such objectives as measuring the effectiveness of alternative advertisements or appeals. In addition, sometimes guidelines that are generally accepted may be mutually inconsistent for a particular questionnaire. For example: sensitive questions often produce better data if placed near the end of a questionnaire; and, it is generally recommended that important questions be placed near the beginning of a questionnaire to ensure obtaining that information even if a breakoff should occur. However, there may be questions that are both sensitive and important, and their placement is not addressed by these guidelines.

For these reasons, guidelines are not always applicable, even in areas where they exist.[2] To construct a questionnaire that causes the fewest problems when used in the field, questionnaire development should be a multistage process during which problems are systematically identified and either eliminated or minimized. The approaches described here can be useful components of this process.

[2]For more extensive discussion of reasons for the inadequacy of proceduralizing guidelines for the design of forms, see Wright (1981) and Duffy (1981). These discussions are also applicable to the design of questionnaires.

Part II

Tools for Developing Questions

This part of the report describes three tools to obtain information that will be useful in the task of actually drafting the questions and assembling them into a questionnaire for a proposed survey: (1) unstructured individual interviewing, (2) qualitative group interviews, and (3) participant observation. In some instances, these same techniques are used later in the process, i.e., during testing or the survey itself, to provide information that will aid in the interpretation of the test or survey results. However, the emphasis in this section is on the aspects of these techniques that contribute to the initial development of the content of a questionnaire.

It is assumed here that a determination has been made that certain information is needed to address a problem and that a survey is the best way to provide this information. Obviously, this determination should be made only after it has been ascertained that the information is not already available elsewhere (e.g., from existing survey data, other records or research studies) or more easily obtained by another method such as the use of administrative records. To make a determination, the problem should be clearly stated, including its possible causes and the potential solutions. The temptation to start drafting a questionnaire before this is done should be avoided. Without a thorough analysis of the problem, the resulting survey may not provide the right information or enough information to solve the problem. The objectives of the survey, including what data should be

collected and how it will be used, need to be directly related to the solution of the problem.

It may be possible to examine a problem and develop survey objectives by researching literature on the topic and through discussions with experts in the problem area. However, information or experts may not be available, particularly if a survey on the topic has never been conducted before. In that case, the techniques described here may be useful in obtaining the necessary background information. Since, in our information-rich society, the necessary information is usually available from other sources, these techniques are not used as frequently to develop questionnaires as some of the other methods described in this report. However, they are included here to ensure that questionnaire designers are aware of their possible uses. Each of these techniques is briefly described below.

Unstructured individual interviewing, described in Chapter 2, is a discussion of the proposed survey topics between an individual member of the group to be surveyed and the questionnaire designer. It is guided by a topic outline rather than a set of specific questions. This technique is used primarily to gain insights into the best way to structure the questionnaire.

Qualitative group interviews, the subject of Chapter 3, are informal discussions of selected topics between participants chosen from the population of interest and someone who is knowledgeable about group interviewing techniques and the purpose of the survey. The information from qualitative group interview sessions can aid in developing the conceptual framework and data specifications for a statistical survey and evaluating draft questionnaires. Qualitative group interviews are occasionally used after a survey has been conducted to help the analysts interpret the data.

The last of these three techniques is participant observation research, described in Chapter 4. While it is not used frequently in designing questionnaires, it can be particularly useful when a survey is to be conducted among people whose language, values, or experiences are very different from those of the questionnaire designers. Information obtained through participant observation can be used to ensure that the content of the questionnaire will provide enough information to satisfy the survey's objectives and to help phrase questions that can be understood by the respondents. The information can also be used to help understand the meaning of respondents' answers to survey questions.

Chapter 2
Unstructured Individual Interviewing

I. INTRODUCTION

Sometimes a questionnaire designer is required to develop a questionnaire on a topic which (s)he knows little about, and about which little information related to questionnaire design exists from previous surveys. In this situation, the development of a questionnaire can benefit from the use of unstructured interviews with members of the intended respondent universe. The term "unstructured interview" is used here to describe a discussion of the proposed survey topics between a member of the target survey population and the questionnaire designer.[1] The discussion is guided by a topic outline rather than a set of specific questions. When sufficient numbers of such interviews are conducted with respondents who are fairly representative of the target population, the technique can provide ideas and insights about how best to structure the questionnaire before the first draft is written.

It is a particularly valuable technique when there are many divergent interests in a survey. When there is more than one sponsor, initial disagreement can exist about what kinds of information can and should be obtained. This technique transfers the questionnaire design decisions from dependence on the tastes or preferences of the survey sponsors to reliance on the results of the field processes employed.

Several of the questionnaire design issues described in Chapter 1 can be addressed by using this technique. The specific uses of unstructured interviewing include the following: (1) Topics previously thought to be important for inclusion can be discarded as unnecessary or irrelevant, and topics which had previously been neglected can be identified as important in fulfilling the objectives of the survey. (2) A determination can be made as to whether the information requested in the survey is readily available to respondents and whether particular kinds of questions can be asked. (3) An evaluation can be made of which topics might be especially sensitive to respondents.

---

[1]This technique was initiated and has been used extensively by survey researchers in England. Researchers in this country were introduced to the technique by Jean Atkinson of the Social Surveys Division in England; it is described in Atkinson (1968) and Hoinville et al. (1978).

Principal Contributor: Anitra Rustemeyer Streett

(4) Assistance can be provided to determine how to phrase particular questions so that the vocabulary is familiar to respondents and the words mean the same thing to all respondents. (5) Decisions can be made concerning the preferability of open- vs. closed-ended questions to obtain particular types of information, and a range of answer categories for closed-ended questions can be specified. (6) An identification can be made concerning who in a household or business is in a position to respond most accurately to questions on the survey topics and, therefore, would make the best respondent. (7) Suggestions can be made concerning the optimal order of questions or survey topics. (8) Insights about which aspect of a topic appeals most to people may be used to determine the best way to approach respondents in order to encourage their cooperation.

II. METHOD

A. Personnel and Skill Requirements

A key concept in the successful use of unstructured interviewing is flexibility. The questionnaire designer functions as a researcher during this process, and must keep the objectives of the study firmly in mind while dismissing any fixed ideas about how to structure the questionnaire.

Best results are achieved when several people, including one who serves as a team leader/questionnaire designer, work together as a team. The team should include interviewers as well as data processing and subject matter specialists. This allows diverse ideas and insights to be used in the refinement of the survey instrument.

Persons selected to conduct unstructured interviews should be experienced interviewers and be capable of understanding the broad perspective of the research project for which the questionnaire will be designed. This type of interviewing requires skills different from those for structured interviewing (i.e., interviewing in which questions are read verbatim from a questionnaire), and only some interviewers on a regular field staff are likely to possess those skills.

Interviewers selected for this type of assignment should feel comfortable "thinking on their feet" as they will not have a questionnaire script to use as a crutch; if they are easily flustered or confused, they give respondents the impression that they are incompetent or that the study is unimportant. Members of the interviewing team need sufficient experience in unstructured interviewing to be sensitive to the effects of wording changes and to recognize responses that indicate potential problems with question wording or order. In addition, interviewers should be able to tolerate long pauses while the respondent thinks or looks for answers, have the ability to probe nondirectively to get the respondent's ideas, and have a thorough understanding of potential problems in questionnaire design which can affect the achievement of the survey objectives.

Some of the interviewers should be highly knowledgeable and skilled in structured interviewing techniques. This allows informed judgments to be made concerning the kinds of things which can be asked in a closed-ended format

and what topics respondents can be expected to respond to within the discipline of a structured interview.

Unstructured interviewing is actually a combined data collection and analysis process. In addition to the interviewing skill necessary for successful results, a "coder" who is capable of making independent judgments is an essential part of the process. This person should be able to analyze and tabulate results of the previous day's work while the interviewers are in the field conducting additional interviews and then meet with them to explain how and where they are failing to meet survey objectives. The simultaneous conduct of these two tasks speeds up the questionnaire refinement process.

Finally, sponsors or subject matter specialists can provide valuable insights in the frequent meetings held to charter the course of the work.

B. Selection of Respondents

Respondent selection for unstructured interviews generally involves purposive rather than systematic sampling. Although rigorous scientific selection procedures are not necessary, respondents should be members of the population to be surveyed and should be fairly representative of that population.

The characteristics of people asked to be respondents for unstructured interviews may depend on the survey topic. For example, in developing a questionnaire dealing with saving habits to be administered to a national cross-sectional sample, the initial round of developmental work may include interviews with people from a variety of demographic population subgroups. During additional interviews, however, different classifications of saving habits may emerge, and it may be necessary to locate and interview persons who are members of specific categories. Thus, the "sampling" of respondents is an iterative process, too--as is the questioning of those respondents.

Respondents may be located by contacting community or business organizations, or by selecting residential areas.

C. Preparation

Before embarking on this phase of a questionnaire design project, the team leader should become familiar with the objectives of the study and make a list of the data elements which are considered necessary to meet those objectives. These data elements include topics and concepts which are particularly vital to the quality of the study, or are otherwise thought to be related to the survey objectives. Prior to the first discussion with a respondent, the team leader should prepare some alternative orders in which the topics might be discussed, as well as any specific words or phrases to be used in relation to any particular topic.

The team leader's next task is to develop the work sheets to be used by the interviewers and coders. Those serving as interviewers should review the materials and meet with the team leader to discuss study concepts and objectives. The interviewers need to be provided guidance, so they will not go beyond the scope of the project.

D. Operation

Interviewers may begin each interview by explaining that they are working on a very early phase of preparing a new survey. They should emphasize the reasons for and importance of talking to people before a questionnaire is prepared.

During this type of interviewing, the interviewer should follow up on answers or comments that seem to have a bearing on how a concept is interpreted by the respondent or how a sequence of questions should be ordered. The interview should have a conversational flavor rather than the question-and-answer format of a formal interview. Inteviewers should understand that their objective is not to collect data in the usual sense--rather, it is to become aware of any difficulties that are likely to arise when the survey is conducted.

Throughout this process, extensive note-taking is valuable, so that insights gained during an interview are not lost or confused with other interviews. Verbatim recording, by shorthand or speedwriting, is ideal for this purpose; however, such a skill is not within the repertoire of every skilled interviewer. Even very abbreviated note-taking can make it possible for an interviewer to return to statements made earlier by the respondent. Following up immediately on some statements could take the interviewer off the topic being pursued; but "passing remarks" and apparently contradictory statements by the respondent can provide additional insights on how to phrase survey questions.

During each unstructured interview, the interviewer should record how each key inquiry was phrased, as well as the wording used by the respondent in answering the question. (Since interviewers often respond to the answers of respondents with idiosyncratic or instinctive phrases of their own, it may be more difficult to remember their own words than those of the respondents.) Notes should be made (during or immediately after the interview) concerning the ordering of the inquiries (if different from the outline), how one topic relates to the next, if and how they overlap, what effect topic order has on the flow of the interview, the respondent's reaction to specific questions of interest, and the apparent level of difficulty of the inquiry for the respondent.

Tape recording, with the respondent's permission, can be useful as long as time is available to listen to the tapes and extract information from them. Ideally, the team leader, team members who are conducting the unstructured interviews, and coders should meet frequently to discuss what they have learned to date. The reason for frequent meetings is to allow all interviewers to gain insights from the experiences of the others and to help one another interpret respondents' comments. Under the guidance of the team leader, changes to the topic outline should be made to refine ideas on how to present topics and sections of the questionnaire, and the order in which to present them. As experience using the topic outline is gained, interviewers will develop their preferred question wording for topics. They should exchange those wordings during their meetings and then try the wordings used by others in successive iterations of interviewing.

The input of the coder is beneficial in noting ambiguities or superficiality in the responses obtained in previous interviews which require further clarification before the response can be coded. Also, the relative frequency of responses to open-ended questions, the range of conditions imposed by respondents on their answers (e.g., "it depends on ..."), and potential response sets can be obtained from the coders' tallies. The coders' analyses and the interviewers' annotated transcripts are discussed among team members, patterns are identified, and suggestions are made concerning potential question formatting, sequencing, etc.

No set number of completed unstructured interviews or days of unstructured interviewing can guarantee a good questionnaire. Perhaps the best indicator that enough unstructured interviewing has been done is the lack of new insights and ideas on question wording and order by team members. The responsible researcher (i.e., the team leader) must judge whether the team has fulfilled its mission, and when the process of putting together the first draft of the questionnaire should be undertaken.

E. Time Considerations

The process outlined here may take longer to complete than drafting a questionnaire without any field work. On the other hand, when the questionnaire is drafted after these procedures have been followed, it is likely to require far less modification; therefore, time required for unstructured interviewing may be wholly or partly recovered later. The exact amount of time involved depends on the number of people who are available to conduct interviews, the number of interviews completed daily by each interviewer, and the iterations of the topic outline, question wordings, etc., required before members of the questionnaire design team are confident to construct a questionnaire.

In general, when the use of unstructured interviewing is incorporated into the development process, 2 to 6 weeks should be allowed in the time schedule. This includes the preparation time for the team leader as well as the interviewing time itself. It does not include completion of the initial questionnaire draft, which would be required regardless of whether or not this technique is used. However, drafting the questionnaire should be much less time-consuming, because the knowledge gained from the unstructured interviews will clarify concepts and resolve most of the issues that are typically debated; e.g., which words to use and which to avoid, how much detail to request of respondents, and the order in which to present topics.

F. Cost Considerations

The monetary costs associated with the use of unstructured interviews are essentially limited to the salaries of the personnel who are members of the team. Depending on the number of people involved, the number of interviews conducted, and the amount of time spent in analyzing the interviews, these costs could vary considerably. In addition, other expenditures may be necessary for travel if the interviewing site is not located near the duty station of the people working on the project.

One other "cost" should be mentioned here: the burden on the public. Although unstructured interviewing places some response burden on the public,

this investment may be more than repaid later if the unstructured interviewing results in a more efficient questionnaire than would be prepared without this type of field work.

G. Mode of Data Collection

Regardless of whether the final survey will be conducted face-to-face, on the telephone, or by mail, the use of unstructured face-to-face interviewing can provide valuable insights on how people respond to the topics of the survey. Benefits accruing from establishing the relevance of specific topics to the survey objectives, defining key concepts, and identifying words which have similar meaning for all types of respondents will be equally pertinent for surveys conducted through any method.

Some of the other insights gained through use of this technique, such as the specification of question order, may be unique to the mode in which the data are collected. If the final survey is intended to be conducted exclusively on the telephone, unstructured telephone interviewing could conceivably be conducted.

III. EXAMPLE: NATIONAL FIRE SURVEY

In 1973 the Bureau of the Census was asked to determine the incidence and characteristics of household fires in the United States. It was decided that a few "screener" questions should be added to the (monthly) Current Population Survey to determine if a fire had occurred in the household within the preceding few months. If a fire had occurred, a separate questionnaire would be administered to gather more detailed information, including extent of damage, death or injury to household members, and financial loss attributable to the fire.

The study directors and sponsors agreed to unstructured field interviews as a means of drafting a questionnaire, because they needed answers to several questions, including what definition of fire should be used and whether people would call things like the following a fire: a grease fire while cooking, a smoldering mattress caused by a cigarette, a small fire ignited by a child, a fire in an automobile engine, a chimney fire. They also wanted to know if questions about injuries, loss of life, and whether the fire was caused by carelessness were feasible and, if so, how to word them and where to place them in the interview. Another area of uncertainty dealt with economic loss and who paid to restore the damage: did people know the dollar value of the losses due to fire and to what extent were they covered by insurance, other family members, charity, etc.?

Since household fires are fairly rare events in the general population, the households selected for the unstructured interviews were chosen from fire department records so that between one-third and one-half were known to have reported a fire within the preceding 6 to 9 months. (See Chapter 10 for a discussion of using record checks.) The other households were selected because they were within two or three blocks of the households identified in the records of a fire department. Choosing nearby households allowed the interviewers to conduct more interviews with less driving time. Interviewers were not told which households had reported a fire to the fire department.

The team leader was a senior member of the survey methods research group; others on the interviewing team were junior professionals from the research and operations offices who would work on the final survey. The team of five worked singly and in pairs and, with permission of the respondents, tape recorded some interviews.

The team began with a list of topics to be covered and a thorough briefing on and discussion of the survey objectives. They met daily to share with the group what they had learned. After 3 days (approximately 4 interviews per day by each team member), patterns of questioning respondents had developed and these were discussed. Agreement was reached on two draft questionnaires. These draft versions were used by all team members during the next 2 days of interviewing. At the end of 5 days of unstructured interviewing it was fairly easy to draft a questionnaire that could be endorsed by all team members as suitable to meet the study objectives and workable with respondents. A definition of a fire was developed which included short lists of things to include and exclude, based on ambiguous areas encountered during the unstructured interviews. The questionnaire was used in an informal test and was judged to work very well. (See Chapter 5 for a description of the objectives and procedures of informal tests.)

Chapter 3
Qualitative Group Interviews

I. INTRODUCTION

By qualitative group interviews we mean open, informal discussions of selected topics by participants chosen from the population of interest, or a subset of that population, led by someone who is knowledgeable about group interview techniques and the purpose of the discussion. Many other terms are used to describe this approach, such as group depth interviews, intensive interviews, focused discussion groups, and focused group interviews. This approach is similar in some respects to unstructured individual interviewing (discussed in Chapter 2) except that it involves a group of participants. The rationale for conducting qualitative group interviews is that information can be brought out through interaction of the participants which would not surface if each of them were interviewed separately. Qualitative group interviews allow closer contact between researchers and respondents than is normally possible in large-scale traditional survey research approaches and permit flexible exploration of research issues from the respondents' points of view.

Qualitative group interviews are an appropriate vehicle for developing insights and hypotheses and for exploring the range of pertinent attitudes, opinions, concerns, experiences, and suggestions of the participants. They can be a helpful preliminary step in developing the conceptual framework, data specifications and question wording or evaluating draft questionnaires for a quantitative survey which will use structured questionnaires among a representative sample of respondents. In the example provided at the end of this chapter, the technique was used to evaluate proposed revisions to an existing administrative form. Qualitative group interviews are also sometimes undertaken solely to provide general information or to help determine whether quantitative research on a subject is feasible; occasionally, they are employed after a survey has been conducted to help the analysts interpret the data that were collected.

II. METHOD

A. Personnel and Skill Requirements

Qualitative group interviews require the services of personnel with specific types of expertise. A discussion leader should be skilled in guiding the group interview within the topical area limits, covering all germane areas,

Principal Contributor: Margaret Weidenhamer

probing for the meaning of comments which are not self-explanatory, yet remaining as unobtrusive as possible to avoid "leading" participants. It is his or her function to initiate discussions among group members and encourage all to join in the discussion, to subtly direct the discussion to the pertinent issues, to prevent domination of the group by any of the participants, and to bring the discussion back into focus whenever it digresses into irrelevant areas. More than one discussion leader may be used depending on the number of groups to be interviewed.

The discussion leaders usually summarize the results of the discussions. For this part of the task, analytical skills are required.

B. Selection of Respondents

The participants in qualitative group interviews are members of the population of potential respondents to the planned survey, but they may not be representative of that group. Usually, a relatively homogeneous group of people, such as middle income city dwellers or suburban homemakers with school children, are invited to participate in a given session. They are chosen by whatever nonprobability techniques may be convenient.

During this phase of survey development, a number of group sessions are generally conducted. The total number of sessions conducted for a particular survey varies considerably. Normally at least four to six group interviews are conducted, and many more may be desirable for complex projects. When multiple sessions are held, different types of people in the target population may be recruited for different sessions. For example, in the development of a national survey on some topic, some qualitative group interviews may be conducted with young black males, others with middle-aged white females. It is also advisable to conduct sessions in several different geographic locations to reduce regional biases.

Participants are usually paid a set fee or a donation is made to an organization of their choice in recognition of the time they spend and the incidental expenses they incur in attending the session. (The need for payment as an inducement to participate must be satisfactorily demonstrated to obtain OMB approval for compensating respondents in Federally-funded surveys.)

C. Preparation

An outline of topics to be covered is usually prepared in advance; it is likely to start with fairly general topics and gradually focus more on details of the subject matter of interest. The outline may be revised between sessions, as the scope of the research becomes more focused.

D. Operation

Generally qualitative group interviews are held in a central location which is convenient for participants, and are scheduled to run for about 2 hours. From 8 to 12 persons are suggested for participation in a given session; some additional invitations may be extended to allow for attrition. When conducting qualitative group interviews, the discussion leader's first task is to create an informal setting that encourages a frank, open discussion

among all the participants and to start the conversation off in the right direction. The approach used must not be so structured that the participants cannot engage in spontaneous discussions which would shed light on their views--particularly views which may not have been anticipated in the topic outline. The outline is usually used as a guide by the leader, but (s)he should allow the discussion to follow its natural course, unless it strays too far from the purpose of the session.

Projective techniques and self-administered forms may be used during the session, and questionnaires or other exhibits may be displayed. Sometimes a series of two or more sessions is held with the same participants, perhaps with a homework assignment in-between. Also, follow-up individual interviews may be conducted with participants. The sessions are usually tape-recorded, and occasionally video-taped, to permit detailed study of the contents.

A good deal of subjective judgment is involved in the analysis of such sessions, and the results must be interpreted with caution. The reports are often written by the discussion leaders who conducted the group interviews. Results should be presented in narrative form, not in terms of proportions or percentages, to avoid suggesting spurious interpretations.

E. Time Considerations

Group interview sessions can be planned, conducted, and analyzed in approximately 2 to 4 months. The time will vary depending on the number of sessions conducted and their locations.

F. Cost Considerations

Qualitative group interviews are a relatively inexpensive way to collect background information for use in developing a questionnaire. The major expenses are the salaries of the recruiter(s) and discussion leader(s)/analyst(s) and the fees paid to participants or donations made on their behalf. Travel expenses and belated costs such as rental of conference rooms and taping equipment may increase the cost considerably if multiple sessions in various geographic locations are held.

G. Mode of Data Collection

A survey employing any mode of data collection--face-to-face, telephone, or self-administered--can potentially be improved through the use of qualitative group interviews in the early stages of questionnaire development.

III. EXAMPLE: EVALUATION OF A PROPOSED REVISION OF AN APPLICATION FORM

The Social Security Administration of the U.S. Department of Health and Human Services employed group interviews to assess a proposed revision of the application form for a social security number.[1] The proposed form contained

---

[1] The information presented here is selected from reports by Bayton (1978) and Scherr (1980).

three pages of instructions and other relevant information, and a one-page application. The application part of the form is shown in figure 1.

Ten group interviews were conducted in a 1-month period in the spring of 1978. They are described below.

Location	Respondents	Number of-- Groups	Number of-- Respondents
Washington, D.C.	Male teenagers; black; low socio-economic status (SES*)	2	10, 12
Washington, D.C.	Female teenagers; black; low SES	1	7
Glen Burnie, Md.	Male and female teenagers; white; lower and middle class SES	2	14, 14
Glen Burnie, Md.	Male and female adults; black and white; lower and middle class SES	1	15
Los Angeles, Calif.	Spanish-language background male and female adults; low SES	2	11, 12
Los Angeles, Calif.	Spanish-language backgrounds; male and female teenagers; two Asians in one of the groups; low and middle class SES	2	14, 15

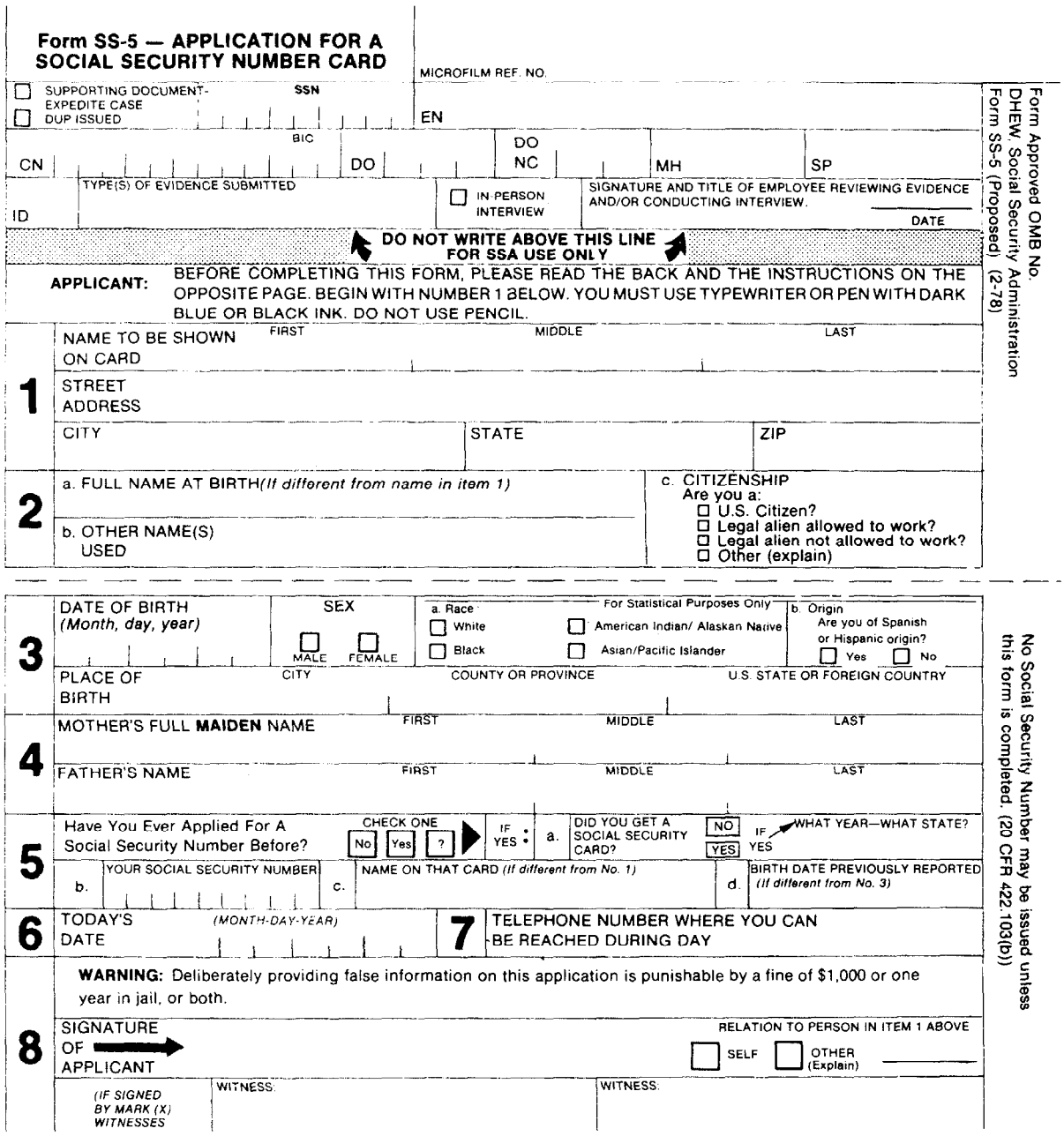
*NOTE: SES is used here as a proxy for expected level of functional literacy.

Each session was tape-recorded and lasted approximately 1-1/2 hours. Each adult respondent was paid $15; each teenage respondent was paid $10. The teenage group sessions did not last as long as those with adults. One or two researchers involved with the project observed each of the sessions; other Social Security Administration staff members also attended some of the sessions.

The group session topic outline followed this sequence:

1. Introduction--purpose of the project.
2. Why should a person apply for a social security number?
3. When should a person apply for a social security number?
4. How can a person go about applying for a social security number?
5. What information does Social Security want from applicants?
6. What documents are needed and why?

Figure 1. Proposed SS-5 Form (to be completed by applicant)



Form SS-5 — APPLICATION FOR A SOCIAL SECURITY NUMBER CARD

MICROFILM REF. NO.

☐ SUPPORTING DOCUMENT-EXPEDITE CASE
☐ DUP ISSUED

SSN

EN

BIC

CN | DO | DO NC | MH | SP

ID | TYPE(S) OF EVIDENCE SUBMITTED | ☐ IN-PERSON INTERVIEW | SIGNATURE AND TITLE OF EMPLOYEE REVIEWING EVIDENCE AND/OR CONDUCTING INTERVIEW. | DATE

DO NOT WRITE ABOVE THIS LINE
FOR SSA USE ONLY

Form Approved OMB No.
DHEW, Social Security Administration
Form SS-5 (Proposed) (2-78)

APPLICANT: BEFORE COMPLETING THIS FORM, PLEASE READ THE BACK AND THE INSTRUCTIONS ON THE OPPOSITE PAGE. BEGIN WITH NUMBER 1 BELOW. YOU MUST USE TYPEWRITER OR PEN WITH DARK BLUE OR BLACK INK. DO NOT USE PENCIL.

1
NAME TO BE SHOWN ON CARD — FIRST | MIDDLE | LAST
STREET ADDRESS
CITY | STATE | ZIP

2
a. FULL NAME AT BIRTH *(If different from name in item 1)*
b. OTHER NAME(S) USED
c. CITIZENSHIP
Are you a:
☐ U.S. Citizen?
☐ Legal alien allowed to work?
☐ Legal alien not allowed to work?
☐ Other (explain)

3
DATE OF BIRTH *(Month, day, year)*
SEX ☐ MALE ☐ FEMALE
a. Race — For Statistical Purposes Only
☐ White ☐ American Indian/ Alaskan Native
☐ Black ☐ Asian/Pacific Islander
b. Origin
Are you of Spanish or Hispanic origin?
☐ Yes ☐ No
PLACE OF BIRTH — CITY | COUNTY OR PROVINCE | U.S. STATE OR FOREIGN COUNTRY

4
MOTHER'S FULL **MAIDEN** NAME — FIRST | MIDDLE | LAST
FATHER'S NAME — FIRST | MIDDLE | LAST

5
Have You Ever Applied For A Social Security Number Before? CHECK ONE ☐ No ☐ Yes ☐ ? IF YES • a. DID YOU GET A SOCIAL SECURITY CARD? NO / YES IF YES WHAT YEAR—WHAT STATE?
b. YOUR SOCIAL SECURITY NUMBER | c. NAME ON THAT CARD *(If different from No. 1)* | d. BIRTH DATE PREVIOUSLY REPORTED *(If different from No. 3)*

6
TODAY'S DATE *(MONTH-DAY-YEAR)*

7
TELEPHONE NUMBER WHERE YOU CAN BE REACHED DURING DAY

WARNING: Deliberately providing false information on this application is punishable by a fine of $1,000 or one year in jail, or both.

8
SIGNATURE OF APPLICANT
RELATION TO PERSON IN ITEM 1 ABOVE
☐ SELF ☐ OTHER (Explain)

(IF SIGNED BY MARK (X) WITNESSES)
WITNESS: | WITNESS:

No Social Security Number may be issued unless this form is completed. (20 CFR 422.103(b))

7. Completion of the proposed form. Respondents filled out the form as though applying for a social security number. They were requested not to interrupt for questions or comments but to wait until the entire group had finished.

8. Open-ended inquiry. Questions or comments initiated by respondents after all members of the group had finished completing the form.

9. Directed inquiry. Item-by-item probing by the discussion leader of respondents' reactions to the form.

Some of the findings from this study were--

The most consistently salient problem was with the race/origin part of item 3. In most groups, this was the first matter raised by the participants. The problems included not understanding what "origin" meant and what "Hispanic" meant. Another salient problem had to do with item 2b--Other name(s) used. Respondents who raised this issue asked whether nicknames were to be included. Father's name (item 4) was mentioned as a problem by some; did this mean "real" father or stepfather? The instructions for item 4 mention stepfather, but it is not clear as to which should be used, if there is a choice. Among teenagers, item 7--Telephone number where you can be reached during the day--was a problem. Did this mean that they should give the number of the school being attended?

Upon receiving the proposed form to fill out, only in rare instances did the respondents read the instructions on the page facing the application or turn the form over and read what was on the back, despite the request to do so which was printed on the application. The more usual use of the instructions came when an individual stopped working on the application and referred to the instructions for an item on an as-needed basis. When asked why they did not read the instructions initially, typical comments were: "I've filled out application forms before;" "I didn't read them; it looked easy," and "If you look over it and you understand most of the questions, you don't need to read the directions..."

When the discussion leader went through the entire form section-by-section, additional difficulties surfaced. For example, the statement "For statistical purposes only" appears over the part of item 3 that contains the race/origin information; many of the teenagers and foreign-language background respondents did not understand the intended meaning of the phrase. Some of the respondents associated this term with the Government keeping a "record." Others said the term referred to the fact that the information asked would not be used in relation to particular individuals. The problem with item 5, Have you ever applied for a social security number before?, was the interpretation to be placed upon the word "you." Several teenagers reported that their mothers had obtained social security numbers for them when they were much younger. If the "you" were to be taken literally, these respondents would check "No." The instruction for this item does not address this problem.

The Social Security Administration used the information obtained through the qualitative group interviews in developing further revisions of the proposed new application package. Controlled testing of specific alternatives was then conducted with larger samples of actual applicants under operating conditions.

Chapter 4
Participant Observation

I. INTRODUCTION

Participant observation research techniques have traditionally been used by anthropologists to study other cultures. By living among people and studying them as unobtrusively as possible, anthropologists have learned much about societies that were relatively unknown. Participant observation research can also be used as a preliminary stage in the design of certain questionnaires. It can be particularly useful in planning a survey among people whose language, values, or experiences are very different from those of the questionnaire designers, or about whom very little is known.

Understanding the culture of potential respondents through participant observation research contributes to questionnaire design in several important ways. First, it increases the likelihood that meaningful inferences can be drawn from respondents' answers. A questionnaire designer who is familiar with the values and experiences of a population is in a better position to write questions which make sense to respondents and to which they will respond more willingly.

In addition, a researcher who is familiar with the population suggested for study knows how to contact individuals with a greater probability of being knowledgeable about the survey topic. Participant observation research helps a questionnaire designer distinguish significant categories of people within the respondent community, and helps in identifying characteristics that may be associated with response. If participant observation indicates systematic differences among age groups, or occupations, or backgrounds in the way topics are conceptualized, the designer may find that a complex questionnaire design is necessary with several paths within a single questionnaire, or even use of multiple questionnaires. This may be necessary to ensure meaningful questions to respondents from different age groups, occupations, or backgrounds.

Surveys of the population of the United States run into frequent problems with respondents who have difficulty understanding questions written in English. In November 1979, the Current Population Survey estimated that nearly 18 million Americans (almost 8 percent of the population) used a lan-language other than English at home. In addition to the potential language problems in a national sample, there are many subpopulations where a much larger proportion of respondents need special questionnaire designs. For

Principal Contributor: Catherine J. Baca

example, the population of Puerto Rico is routinely the subject of Federal surveys dealing with employment, the labor force, or food assistance. Recent Asian and Caribbean immigrant populations have been asked to respond to Federal surveys about immigration, literacy, and public assistance. These respondents, along with many employment and income groups that use special vocabularies or share distinct cultural outlooks, require questionnaires that are written specifically for them.

II. METHOD

A. Personnel and Skill Requirements

There are three different ways in which participant observers can take part in the design of a questionnaire. First, when little or nothing is known about the respondent universe, participant observation data can be collected by a field researcher selected for this purpose. Such a person (or persons) might be recruited through university graduate departments of anthropology or through national professional organizations[1] that maintain records of their members' professional experience and research skills.

Second, fieldworkers who have previously conducted participant observation research among the potential respondents can be involved in the questionnaire drafting phase of survey development (either on a full-time or consulting basis). In this way, insights into potential difficulties in respondent understanding and/or interpretation of the questions, respondent perceptions of the subject matter, etc., can be incorporated into the survey instrument. The example presented at the end of this chapter describes this use of the technique.

Third, published data based on participant observation research can be used by questionnaire designers for certain projects. It is not likely that published monographs can be used to find solutions for specific questionnaire issues, except for very large populations which have been the subject of extensive research. But some combination of the second and third techniques will provide needed design assistance for most smaller populations.

B. Selection of Respondents

It is simple to state that the design or purpose of a survey dictates the selection of a respondent universe for participant observation. The task is far more difficult in practice. There is a complex literature on how to define an appropriate community for specific ethnographic research goals. If a survey is contemplated in a residential community, the universe for participant observation is easy to define geographically. But in the United States it is more common to conduct a survey among respondents defined by some characteristic besides residence. Again, the general rule is that the respondent universe for participant observation is bounded by the goals of the survey. In practice, the purpose of participant observation research

---

[1] These include the American Anthropological Association, the Society for Applied Anthropology, and the Washington Association of Professional Anthropologists. Each of these organizations is headquartered in Washington, D.C.

may actually be to learn the boundaries and significant characteristics of the respondent universe. Many times it is up to the participant observer to discover who the potential respondents should be, if the goals of a survey are to be achieved.

The role of individual respondents is discussed in more detail in part D, below. But it should be noted here that the results of participant observation depend upon the representativeness of the informants. Since much of the information is collected from a limited number of people, it is possible to make errors related to population variability. To avoid this, participant observers should make an effort to ensure that they observe and interview a variety of people. Whenever possible, more than one researcher should be involved in conducting the fieldwork, to reduce the likelihood of significant errors.

C. Preparation

The selection of participant observation field researchers can be done in consultation with one of the professional sources described under II-A, above, or through examination of published literature on the respondent community.

Some researchers require formal introduction to "the field." This could be through personal letters of introduction to members of the respondent community or through temporary association with an institution with which respondents are connected in some way (for example, as employees or clients). In other cases, participant observation begins simply when the fieldworker travels to the location where respondents are to be found.

D. Operation

Participant observation is distinguished by four characteristics: Use of the respondents' own language; residence or participation in the respondents' community; key informants; and unstructured interviews.

1. Using the Respondents' Language

The importance of conducting research in the respondents' own language may be easier to understand if "language" is thought of in the widest sense. A difference in "language" may be a regional dialect or a professional jargon. Two groups who speak the same "language", such as English, may have regional or cultural differences that cause them to infer very different meanings from the same words or arrangements of words.

First, using a translator or bilingual interviewers will not solve the fundamental problem of assigning valid meanings to the answers of non-English-speaking respondents. If respondents make the translations needed for answering questions, their decisions about what to include or exclude in the meaning of words may be far different from what the designer intended. It is the designer's responsibility to ensure that there will be no differences between what (s)he means and what respondents mean when each uses the questionnaire.

Second, the period spent in learning the respondents' language has a value of its own in the research process. The participant observer's obvious effort to learn the local language makes him/her more acceptable to potential respondents and reduces the disruption that an outside observer causes. As a result, valid observations can be made sooner than with more intrusive techniques. In addition, because language embodies culture, an observer learns much more than language. The frequency with which certain words, phrases, and concepts are used has often been a vital clue to researchers. In sum, the effort to learn a local language improves a researcher's efficiency and ensures that (s)he can recognize any potential failure to communicate.

2. Living Among Respondents

A participant observer can gather information about a community in a variety of ways. The traditional approach involves living among the people being studied. However, participant observer research methods are also used to learn about groups that come together only at limited times or places, for example, ethnic groups or employee groups such as nurses. To study nonresidential, scattered communities such as these, participant observers spend as much time as they can with their subjects, over weeks or months, whenever the group is together.

By spending relatively long periods of time among respondents, a participant observer accomplishes three things that cannot be accomplished as efficiently by any other means. First, there is an opportunity to study the variety of activities and people in the community without prejudging which is most significant. Second, a participant observer learns about the values of the community because to some extent the members' experiences are shared with them. Third, by acknowledging the research role and by seeking respondents' opinions, a participant observer earns the trust of the respondents. As a result, many respondents develop an interest in the research and even look for ways to assist.

3. Key Informants

Participant observers find that much of their information is collected from key informants. Key informants are individuals who are willing to talk at length with the researcher, or who serve as an entree to many further contacts, or who reveal extraordinary knowledge about some topic. They provide richly detailed information about people and institutions. For example, elderly people are sometimes key sources of historical or geneological data.

A researcher who uses key informants does run the risk of collecting data from an individual who is not representative or who tries to mislead. These risks can be minimized through checking what is learned with a variety of other informants and through observation. There is also no reason why participant observation research cannot incorporate the principle of randomization of informants at some stage, as a check on key informants, or to counteract the fieldworker's own potential bias. But at other stages of the research, such as entry to the field, participant observation succeeds precisely because informants are allowed to volunteer.

4. Unstructured Interviews

In addition to residence among the respondents, the participant aspect of the research involves unstructured interviews. This technique, which is described in Chapter 2, is particularly suited to the study of groups about which little is initially known. Unstructured interviews permit hypotheses about survey content and questionnaire construction to be tested and rejected very quickly. For this reason they are particularly appropriate to the beginning stage of questionnaire design.

In developing a questionnaire for a respondent group that is not well known, however, methods such as unstructured interviews that yield consistent results do not necessarily yield meaningful results. If someone unfamiliar with a culture asks a limited number of questions, (s)he can get consistent responses, yet err in the meaning attributed to the responses. This is so for three reasons. First, interviews are artificial situations in which respondents may tailor answers based on their perceptions of what the questionnaire designer wants to achieve. Second, the questionnaire designer does not necessarily understand patterns of bias among respondents. For example, are there distinctive respondent strata represented? Third, the questionnaire designer cannot easily cross-check the results of these techniques. Results should be compared to responses derived in other situations, at other times, and from other respondents. Without alternative sources of information about the respondent population, interviews do not preclude major errors of interpretation.

Consistency of results from such techniques may mean only that a "structured misunderstanding" is occurring. This phrase has been used recently to describe consistent and self-perpetuating mutual misunderstanding between U.S. census takers and members of a minority subculture (Hainer, 1979). In some cases when dealing with respondents from another culture, failure to communicate is recognizable. But misinterpretations might also go unrecognized. This is similar to a translation problem; if a phrase in language A is translated into language B, the words might make sense without it being in any way the sense intended. If the translation satisfies the expectations of those who speak B, no one will suspect a mistake. Complementary misunderstandings such as those described by Hainer can even permit groups to appear to cooperate. There is no single research technique that will uncover such mutual misunderstanding. But the multiplicity of methods used by a participant observer makes such an occurrence very unlikely.

5. Variations of the Method: How Much Participation?

Participant observation research methods are on a continuum from unobtrusive observation to total immersion in a community as a member or actor. The optimum combination of methods depends on the characteristics of the researcher, the topic being studied, and the characteristics of the research subjects. As an example of the range of personal research styles, consider two studies of social organization among low income urban Black communities in the United States. One participant observer moved her household, including her children, to live among the families she was studying (Stack, 1974). Another participant observer was able to cultivate personal relationships in a similar community without leaving his own home. Every day he visited the

neighborhood he was studying and spent the day with his informants (Liebow, 1967).

At the "observation" extreme of the continuum are studies of groups that could not be conducted by a resident fieldworker. If the research subjects do not live together, for example, it is impossible for a researcher to live among them. Participant observers have studied longshoremen, vagrants, and ethnic communities, for example, by visiting the subjects at the times when and places where they come together. Studies which focus on institutions, such as hospitals, factories or schools, are conducted primarily on site, at the times when respondents are willing to talk to a researcher.

At the opposite end of the continuum are data collected while the researcher is a member of the subject community. The researcher might join a community to collect data, or might analyze an organization or group to which (s)he already belongs. Clearly this end of the continuum gives a researcher maximum access to insiders' values and behavior, but it is not always preferable. Not only does it create ethical dilemmas (i.e., subjects may not be aware of the researcher's intent), it often reduces the observer's capacity to interpret the observations. An insider lacks the outsider's awareness of alternatives, which is the first step to analyzing existing cultural elements. For the purposes of designing a questionnaire, participant observation research would generally tend toward the formal observer end of the continuum, as opposed to the member/participant end.

Participant observers can present themselves in a variety of roles, ranging from the potentially unsettling identity of an outsider with no familiar attributes, to roles known to respondents such as student, government agent, adopted family member, etc. The purpose of selecting a role from among those available (or changing roles) is to minimize the obtrusiveness of the participant observer's presence while maximizing the likelihood of situations that provide useful observations. An experienced researcher balances obtrusiveness and its potential adverse effect on data quality against the benefits of taking active steps to elicit certain kinds of response.

6. Variations of the Method: How Much Fieldwork?

There are at least two kinds of research questions that can only be answered by spending a relatively long period in the field. The first kind deals with sensitive topics, information that people do not want to reveal. Informants who cannot expect anonymity will only discuss these topics when they trust the researcher, and that trust is developed gradually. The second kind of research question that may require relatively prolonged fieldwork deals with matters of which the respondents are unaware. People are seldom able to answer questions accurately about the relationships between variables in their own society. The problem is made more difficult when generalizations must be made about another society. Observations over time, however, are likely to provide a participant observer with hypotheses about the magnitude and direction of relationships between variables which can be tested through survey research.

If a questionnaire deals with topics which most members of a community are familiar with, and willing to talk about, then the questionnaire designer's

job is relatively easy. In these cases, a questionnaire might be drafted after unstructured interviews are conducted and agreement is reached among most respondents as to the identity and meaning of important topics and concepts. If, on the other hand, a survey will deal with a topic that people are reluctant to talk about, or it is intended as a measure of variables of which respondents have only indirect knowledge, then the participant observation phase of the questionnaire design process is likely to be longer.

E. Time Considerations

Incorporating a participant observation research program into the development of a survey questionnaire may require a substantial amount of time. The specific duration of the research would vary with the survey topic and the type of respondent. For populations about which little or nothing is known, a year in the field might be necessary to obtain useful results.

There are ways to shorten the time required for field research. A search of existing ethnographic literature may locate reports of previous field research which contain useful background information about the survey topic or the respondent universe. This literature can be used as a substitute for extended participant observation, or as a supplement to it. Another way of using the results of such research is to take more direct advantage of an expert's knowledge of the survey population. This can be accomplished by using consultants with relevant fieldwork experience during the questionnaire design process.

F. Cost Considerations

The cost of conducting participant observation research, consisting as it does of support for one or more researchers to live in the field, is usually a very small part of the cost of developing and conducting any survey with a large sample of respondents. If a survey should require extended original participant observation fieldwork, the total cost would be that of keeping the researcher in the field for about a year. That would include the costs of the researcher's travel and subsistence at the local level; supplies such as paper, pens, maps, film or magnetic tape; and equipment such as a camera, tape recorder, and typewriter. Sometimes a researcher also pays a research assistant a part-time local wage, and sometimes a participant observer needs a supply of such commodities as tobacco, medicine, or food as gifts to informants. The direct costs of participant observation are generally so low that they are outweighed by overhead costs incurred when a researcher is affiliated with an institution such as a university.

The costs of incorporating the results of existing participant observation literature into questionnaire development are even lower; the only costs for this are salaries for personnel involved in locating, reading, and interpreting the reports of previous fieldwork. The cost of employing a knowledgeable researcher who has already studied a potential respondent population is also relatively low, consisting of charges for professional consultation during the questionnaire design process.

G. Mode of Data Collection

Participant observation is an appropriate tool for the development of any type of questionnaire. Regardless of whether the method of administration is by face-to-face interview, telephone interview, or mail questionnaire, the knowledge gained through the use of this technique can improve the quality of the survey data.

III. EXAMPLE: 1980 CENSUS OF POPULATION AND HOUSING

Participant observation research was vital to the design of the 1980 Census of Population and Housing as it was carried out in the Outlying Areas of the Pacific. These areas include American Samoa, Guam, the Northern Marianas and the Trust Territory of the Pacific Islands. The traditions, languages, and environment of these islands are so different from those of the United States that the Census Bureau contracted for an anthropologist to serve as a consultant in the design of the questionnaire and the procedures for the 1980 census of the Pacific Islands.

The anthropologist who served as consultant had spent most of the previous decade becoming familiar with the culture and languages of the Pacific area. He had conducted participant observation fieldwork on two atolls in Micronesia, and in American Samoa. His research had required fluency in several native languages, and he conducted censuses of individual communities and islands for research which included geneological, demographic, and socio-economic analyses of island populations.

The questionnaire and procedures used in the 1980 Census of Population and Housing in the Outlying Areas of the Pacific Islands were modified from those of the 1980 U.S. census in three ways. First, there were a large number of changes which reflected the unique characteristics of the Pacific Islands, including differences in environment, technology, and material culture. These were changes in labelling (of names, definitions, or response categories) which made questions and answers more comprehensible to local respondents.

The second category of changes included questions where the content of a question or answer had to be changed as well as the labels. The data collected in the Pacific were, as a result, not exactly like the data collected in the United States. However, the questions used in the Pacific elicited data that could be used in building inferences comparable to those based on responses from the United States.

The third category of changes consists of questions which were added because of the participant observer's knowledge about the culture of the Pacific Island communities. Some questions were ultimately added to the census of the Pacific Islands because anthropological analysis documented their significance to communities in the Pacific.

A. Category 1, Label Changes

Anyone familiar with the characteristics of the people and environment of the Pacific Islands would point out that many definitions and response categories used in the U.S. census were inappropriate for use in the islands.

The answer categories on ethnicity, for example, had to be modified to match the probable responses in the Pacific Islands. Parallel changes were made to the question on place of birth. The answer categories for these questions were selected to represent the most likely patterns of inter-island migration, given the level of specificity permitted in a census.

Local land tenure patterns are reflected in housing questions H29a and H29b. The question on the value of property (H11) which was used in the United States was only appropriate in Guam. In the other Pacific territories, where traditional land tenure is communal, individuals have no precedent for gauging the value of the property upon which their dwellings are built, so the question covered only the value of the dwelling.

ASK H29a IN AMERICAN SAMOA, COMMONWEALTH OF THE NORTHERN MARIANA ISLANDS, AND THE TRUST TERRITORY OF THE PACIFIC ISLANDS ONLY.

H29a. *If this is a one-family house (or condominium unit) which is owned or is being bought –*
What is the value of this house, that is, how much do you think it would sell for if it were for sale? *Do not include the value of the land.*

ASK H29b IN GUAM ONLY

H29b. *If this is a one-family house (or condominium unit) which is owned or being bought –*
What is the value of this property, that is, how much do you think this property (house and lot or condominium unit) would sell for if it were for sale?

H11. *If you live in a one-family house or a condominium unit which you own or are buying –*
What is the value of this property, that is, how much do you think this property (house and lot or condominium unit) would sell for if it were for sale?

Other examples of relatively simple label changes are found in the Questionnaire Reference Book (QRB) and the enumerator's manual prepared for the Pacific Islands. The instructions for recording respondent names, for example, describe the procedures for dealing with hereditary local titles.

> Samoa: Reference to matai title... when a person uses his title as the last name, the people who "belong" to this title may also take this name.
>
> For example, a person whose real last name is Talofa might report his name as John Samoa (the name of his title), and his children might have either Talofa or Samoa reported as the last name, print the last name as reported. [SIC]

Additions were made to the enumerator's manual and QRB to deal with the special characteristics of housing in the Pacific. In the census of the Pacific, respondents were asked what material was used to build the walls and roof of their dwelling. One of the answer categories added was "thatch,"

which is defined as "palm or pandanus thatch, palm leaves, straw, etc." (QRB, p. 97).

Enumerators in the Pacific were taught to calculate a household's annual fuel costs if respondents said that charcoal was purchased by the bag or kerosene by the can (QRB, pp. 112-113). The participant observer knew such replies would be common in the small islands and atolls that predominate in the Pacific Islands.

Finally, the simpler technology of the Pacific territories is reflected in changes made to questions about kitchen and bathroom facilities. A summary question was used in the United States where complete facilities are virtually taken for granted. But positive responses to a summary question would be so rare in the Pacific Islands that separate questions had to be asked about such facilities as hot and cold running water and bathtubs.

B. Category 2, Content Changes

The definitions of a number of questions were changed so that they would generate data comparable to data collected in the U.S. census. These changes were more subtle than the changes discussed in the preceding section. They were based on the participant observer's knowledge about the meaning cultural traits have for respondents. This knowledge was derived primarily from the participant aspect of research, in which the anthropologist became familiar with what respondents think and feel, the language they use, and the relationships of cultural traits to one another. The meaning that certain cultural traits have for respondents and the relationships between traits were reflected in questions about fertility, migration, language and work asked in the census of the Pacific Islands.

Questions on fertility were redesigned to allow demographers to use the data from the Pacific Islands to make analyses and estimates parallel to those calculated for the United States and other places. Earlier attempts to measure individual fertility in the Pacific territories based on questions used in the U.S. census were complicated because there is a higher rate of adoption among households in certain islands. In addition, indirect measures of fertility were needed because vital registration was incomplete in the Pacific territories. To analyze individual fertility, it was necessary first to match children to their biological mothers, regardless of their current residence (e.g., adoption). In the Pacific census, three questions that had no counterparts in the U. S. census were asked for children: Is the biological mother living in the household? Is she still living? And, if she appears on the questionnaire but the relationship is not acknowledged, what is her person number?

Questions concerning children ever born were also expanded in the census of the Pacific to provide better estimates of fertility and mortality. The participant observer's experience indicated that cultural attitudes toward vital registration of such events as infant mortality and adoption made it necessary to ask these additional questions to make data comparable with the U.S. data. So the instructions pointed out that adopted children were not to be reported among children ever born, and, in addition, women were asked how

many of their children were still living, and if any were born alive in the last 6 months.

Migration was a second subject for which a whole series of questions were modified to improve data from the Pacific. Despite the vast distances between the islands, it was not unusual for a significant number of people to be living (for work, school, or other reasons) far from the island of their birth. There was a traditional pattern of temporary migration for most young men on the islands before European contact. Today, young people of both sexes are encouraged to travel to distant education centers in the Pacific territories, or to the United States, for schooling. As a result, many adults live far from their place of birth.

Migration patterns are significant in the Pacific for many reasons. Perhaps the most critical is that, on small islands, population growth can very quickly get out of balance with limited ecological resources, and the greatest source of population shifts in Pacific Islands in this century has been migration.

In the Pacific territories, migration data are also significant because they are relevant to public policies concerned with education and labor. Programs are limited by the willingness of the population to migrate. Analysts need data to measure the potential effect of these policies; for example, is there resistance to migration? What is the rate? What factors cause return migration? Who migrates, and what happens to those who are left behind?

To answer these questions within the limits of the census, each respondent was asked about place of birth (if it was not the place of enumeration), mother's and father's place of birth, and any lengthy period of residence or activity in the United States. The question on residence 5 years ago which was used in the United States was retained as well.

The participant observer was able to predict that questions about language used in the U.S. census would cause problems in the Pacific. Few native residents of the Pacific Island territories use English as their primary language in the home. For the non-European population in the Pacific territories, native languages (and even multiple native languages) would be reported far more frequently than English as the language used at home. The question used in the U.S. census to measure fluency in English (How well does this person speak English?) would be of little use among a population with a majority of non-English speakers. In the Pacific census, therefore, a question for all respondents was designed to identify actual language practice.

Major changes were also made in the questions dealing with the "work" that adults reported doing in the Pacific. Economic activity in the Pacific Islands is very different from that in the United States. A high proportion of adults in the islands derive support from indigenous noncash-related subsistence activities. This includes producing food or goods for home consumption, with little or nothing exchanged for cash or other goods.

In the Pacific, subsistence activity was provided as an alternative response in the question on activity last week. It was also incorporated into the

series of questions on income, partly as a check to ensure that only activities distinct from the cash economy were being reported. These questions identified persons involved exclusively in subsistence activity and distinguished them from persons in the cash labor force who were not working. They also allowed subsistence activity to be reported as a distinct activity pursued along with participation in the cash economy in some form.

C. Category 3, Questions Significant in the Pacific Territories

The final category includes questions that were unique to the Pacific census. These were included because they dealt with topics that are significant in the Pacific Islands and which merit collection of data in an enterprise as costly as the decennial census. The anthropologist was able to help the Bureau evaluate the relative importance of potential census data to the people of the Pacific Islands.

The simplest example is literacy. Citing patterns of native language use and English fluency, the participant observer documented the need for a question on literacy. His experience suggested that data from this question would be important in analysis of programs related to education, training, and employment. In the enumerator instructions literacy was defined as the ability to read or write a personal letter providing an explanation comprehensible both to native enumerators and respondents.

A second major illustration is found in the example of the questions dealing with migration. Of the 31 population items covered in the Pacific census, 9 were directly related to analysis of migration patterns. Because of the immense significance of migration phenomena to the interpretation of a variety of related social processes in the Pacific territories, as described in part B, a lengthy series of questions on migration was eventually included in the questionnaire.

In conclusion, the census in the Pacific territories differed from the U.S. census in many ways. Some of the differences appeared superficial. Other changes allowed the answers from the Pacific to serve the same analytical purposes as answers from the United States. These changes required familiarity with local culture, including knowledge of native languages and native use of English words and categories. Finally, the most fundamental differences in the Pacific questionnaire are reflected in the topics chosen. The anthropologist, serving as consultant, helped the Bureau select the questions which were most valuable for use within the limits of a decennial census.

Part III

Procedures for Testing the Questionnaire Draft

The three previous chapters identified tools that can be used to obtain background information to assist in developing the first draft of a questionnaire; i.e., before any specific survey questions are written. However, other means are more commonly used to obtain such information. For example, the questionnaire designer can review available literature on the topic and questionnaires from other surveys, if there are any, that also addressed the identified data requirements. If another questionnaire exists, persons involved in that survey, if available, and reports on the results should also be consulted as possible sources for learning more about developing a similar questionnaire. Often, unless one's own research indicates otherwise, specific wording of a question can be adopted from another survey. In addition to having wording that has been "tested," it might allow the data to be compared with another source.

Even if other questionnaires on the proposed subject of the survey do not exist, there are several reference sources the designer might use for guidance in writing questions. Since many household surveys include questions on respondent characteristics for categorization into analytic groupings, several attempts have been made to gain acceptance for standard wording of these types of questions. Two such attempts are Basic Background Items for

U.S. Household Surveys (Social Science Research Council, 1975) and Social Concepts Directory for Statistical Surveys (Statistics Canada, 1980). These reports, or others like them, may be useful in determining how to word questions on age, marital status, education, income, etc. Although there is still some debate on the possibility and desirability of standardizing questions, it is generally agreed that even small differences in the wording of a question may affect the resulting data. In addition, many books have been written on how to design questionnaires. Works such as The Art of Asking Questions (Payne, 1951), Designing Forms for Demographic Surveys (Sirken, 1972) and Asking Questions (Sudman and Bradburn, 1982) are valuable sources of general advice on how to write questions and on other aspects of designing questionnaires.

Finally, before the first attempt is made to draft questions, there are some other basic issues which need to be considered. These include such things as the number of interviews with each respondent (more than one may be necessary), the frequency of the interviews, the data collection mode, and the type of respondent. (See Chapter 1 for further discussion of these issues.) The overall structure of the questionnaire should also be established showing the organization and relationship of the various components, pieces, or sections making up the entire questionnaire. For example, a questionnaire may have separate sections or even physically separate documents for different topics covered in the survey and/or for different persons within the household who are to be interviewed. Once the overall structure of the questionnaire is determined, it can serve as a guide for developing the individual questions.

Writing the questions is a critical step because the results of the survey depend on the answers given to each question. The question wording must be clear and comprehensible to most respondents to minimize biasing of the survey results. In addition to writing the questions, the designer must sequence them in a natural order that will flow smoothly from one topic to another. The flow may be improved by using screening questions and skip patterns. Screening questions are specifically designed to determine whether certain questions should be asked of a particular respondent. For example, respondents might be asked if they have any children before they are asked a series of questions about their children; respondents without children would be "skipped over" (i.e., not asked) these questions. Skip patterns are used in the same way to avoid inapplicable questions depending on the respondent's answer to a previous question.

When the first draft of the questionnaire has been prepared, it should be subjected to extensive review. The reviewers should include the analysts and other staff members working on the survey and, whenever possible, other persons outside the staff who are familiar with the topic of the questionnaire or uses of the data. The review process should ensure that the data requirements or objectives of the survey are being met. The draft can also be administered to friends and/or coworkers to check for problems such as skip pattern errors or awkward wording. Sometimes questions which look good on paper sound stiff or verbose when read aloud. The responses to the draft at this point might indicate how respondents selected for the survey will react to the questions. After considering the comments and suggestions received during the review, another draft of the questionnaire will probably

need to be prepared to incorporate revisions. Several iterations of the questionnaire and review process may be necessary before the designers are satisfied with the product.

At this stage, it is imperative that the draft questionnaire be tested with the population under study. This part of the report discusses various ways of testing the questionnaire under field conditions. Field testing is particularly appropriate for questionnaires administered by interviewers in person or by telephone. It also may be used for self-administered questionnaires which are usually mailed to respondents. Another type of testing which is more useful for self-administered questionnaires is laboratory or classroom testing. In this type of testing, a subjective evaluation is made of the questionnaire under controlled or semicontrolled conditions. This is done by having participants complete the draft questionnaire, in a group setting or individually, and then talk with the questionnaire designer(s) about problems encountered. However, only field testing is covered in this report.

This report divides field testing into two broad categories: informal and formal. The main distinctions between tests in these categories are in the size and the sophistication of their sample design and the completeness of their objectives. Informal testing relies primarily on subjective evaluations of the questionnaire; whereas, formal testing relies on statistical evaluations. As the word "informal" implies, less control is necessary in choosing the sample and conducting the interviews for such testing. The next chapter, Chapter 5, describes informal testing in more detail; formal testing is described in Chapter 6 with emphasis on two variations: pilot studies and split sample tests. These chapters describe the circumstances and factors that should be considered in determining the type of testing to be undertaken in preparation for a survey.

Chapter 5
Informal Testing

I. INTRODUCTION

Once the initial version of the questionnaire has been drafted, several types of field tests can be conducted to refine the questionnaire. One type is the informal test. In this report, informal testing refers to a questionnaire field test involving a relatively small number of interviews in the kind of setting chosen for the final survey (i.e., home, work, etc.) as opposed to a laboratory setting. In this type of testing, the detection and correction of errors or weaknesses in the questionnaire draft depends mainly upon subjective information provided by interviewers and observers. The test is not designed to be evaluated on a rigorous statistical basis.

If a series of tests is planned in the questionnaire development process, an informal test is frequently a first step, with formal tests involving more sophisticated types of evaluation coming later in the refinement process. Or, it may be the last step in the process to ensure that the revisions made as a result of previous formal tests work well together. If time and money permit only a single test, the relative speed and low cost of an informal test (in comparison with a formal test) may make it a logical choice.

In terms of the questionnaire design issues outlined in Chapter 1, informal tests are particularly appropriate and useful in discovering poor question wording or ordering, errors in questionnaire layout or instructions, and negative response effects caused by the length of the interview or a respondent's inability or unwillingness to answer the questions. In addition, they can be used to a lesser extent to assess the feasibility of using a particular concept in a questionnaire, to determine if the questions seem to elicit appropriate responses, and to suggest additional questions or response categories which can be precoded on the questionnaire.

Other relevant objective information which might affect the final questionnaire design can also be obtained in an informal test--e.g., a preliminary indication of the interview length (called respondent burden by OMB), the refusal rate, and field costs.

Principal Contributor: Dawn D. Nelson

II. METHOD

A. Personnel and Skill Requirements

Several types of skills are necessary to conduct an informal test, some of which may be combined in a single person. However, it is usually necessary to have a team of persons or several different groups of people.

If a team of persons is used, someone must coordinate all the activities involved. These include selecting the test site, selecting the sample, selection and training of interviewers, developing the questionnaire to be used, structuring a system to receive feedback about the questionnaire, and setting up a plan to evaluate the questionnaire. (Each of these topics is discussed further in the next section.) Experience with or knowledge of data collection operations is an essential qualification for this person.

Some personnel may also be required to conduct interviews. There are advantages in selecting skilled, experienced interviewers for informal tests. With such interviewers it is more likely that question misunderstandings or difficulties will be due to questionnaire design deficiencies rather than to the interviewer. They also can provide considerable assistance in improving the questionnaire based on their experiences with other surveys. However, there are some disadvantages also; e.g., they may inadvertently cover up questionnaire problems by their own deft handling of a situation, something that a less experienced interviewer in the actual survey may not be capable of doing. Thus, the use of interviewers with varying experience and skill levels may be desirable in an informal test. The interviewers should know how to probe to obtain information that will be useful in refining the questionnaire. All interviewers do not possess these skills and should be trained on them, if necessary.

Another option is for the questionnaire designers and researchers to serve as the interviewers. This ensures that the persons doing the interviewing are thoroughly familiar with the aims and objectives of the test. However, only questionnaire designers who are knowlegeable about interviewing techniques should attempt this; otherwise they could adversely affect the test results. Even if they do not plan to perform this role, such training will make them more sensitive to the problems questionnaires can cause interviewers.

In addition, knowledgeable personnel are required to carry out the evaluation of the test results. Skills involved in this activity include ability to recognize problems during an interview, or in a review of the completed questionnaires or tabulations, and the implications of the test results for the design of the questionnaire. Personnel involved in the evaluation should actively participate in the operational phase of the test.

B. Selection of Respondents

Usually, adequate subjective information can be obtained from 50 to 300 respondents. The respondents generally are selected purposively rather than randomly to achieve the desired objectives of the test. For example, if the survey will be conducted with a general population sample, representatives from a broad range of subpopulations should be included in the informal test.

On the other hand, if the questions being tested are directed at a specific subpopulation, such as food stamp recipients or high income persons, the entire test sample might be composed of representatives of that group to ensure adequate coverage with a small number of interviews. When this is the case, the site selection may depend on the location of the subpopulation or the availability of high quality records for use in selecting a sample. (See Chapter 10 for more information on the use of records.) If no such constraints exist, then convenience and low cost are the chief factors in selecting a location, which frequently results in the selection of a site near the agency headquarters.

C. Preparation

The study design for informal tests is probably more important than the number of interviews in ensuring that the results are useful because subjective evaluations are not always improved by the quantity of observations. However, compared to formal tests or the actual survey, the design of an informal test is usually relatively simple. In planning for one, the following factors should be considered:

1. The Questionnaire Composition

A decision should be made on whether to test the entire questionnaire or only a portion of it. If only one test is planned, it is advisable to use the entire questionnaire since responses can be affected by the presence and order of the questions included in the proposed questionnaire. For this reason, questions borrowed from other surveys should not be omitted from this testing.

When a series of tests is planned, one or more of the informal tests may be devoted to a particular portion of the questionnaire that is expected to be troublesome. In such situations, the section tested might be relevant only for a particular subpopulation and the sample for the test might be limited to that population subgroup as discussed above in section B. At the end of this process, the entire questionnaire should be tested to see how the sections work together.

Another questionnaire choice concerns the possibility of using two or more versions of the question (or answer) wording or order. Although this is perhaps a more common technique in split-sample testing (see Chapter 6), it can be used effectively in an informal test to make a quick comparison of the alternatives.

2. The Interviewing Method

Again, the choice of interviewing procedures is affected by whether or not a series of tests is planned. If the informal test will be the only test, the questionnaire probably should be administered in the same manner selected for the survey (e.g., self-administered, interviewer-administered in person or by telephone, or some combination of these methods). However, as part of a series in which the informal test will be used only for a preliminary indication, a different method may be justified to save time and/or costs. If the interviewing method will be the object of later experimental testing, the

informal test could contribute to the planning for the experiment by using all the proposed methods.

3. The Training of Interviewers (for Interviewer-Administered Questionnaires)

If professional interviewers (as opposed to the designers and researchers) are used to conduct interviews, they should be thoroughly trained on the purpose of the test and the concepts and definitions used in the questionnaire, as well as on the proper way to administer the questionnaire. With a better understanding of the rationale and logic behind the questions, the interviewer should be able to make a more significant contribution to the evaluation. If questionnaire designers and researchers who are inexperienced interviewers do the interviewing, they should have an introduction to general interviewing techniques before beginning their assignment.

4. The Observational Feedback System

The most important element in the design could be the system developed to capture the subjective observations on the performance of the questionnaire in the informal test. There are several ways that this can be accomplished. For example, interviews can be tape-recorded, observers can accompany the interviewers and record information on a specially designed evaluation form, the interviewers can be provided with a similar evaluation form to be filled out, or the interviewers and/or observers can be debriefed following the test. Observers are extremely helpful because they can watch the interaction between the interviewer and respondent to detect problems which might not be apparent to the interviewer. (Chapter 8 of this report contains a discussion of observation as a tool for evaluating questionnaires. Chapter 9 contains a discussion of interviewer debriefing and structured evaluations.)

In addition to these more formal mechanisms, the preliminary nature of an informal test allows interviewers and/or observers to initiate conversations with respondents at the conclusion of the interview. In this way, a respondent's impressions about the meaning of certain questions or concepts can be clarified, and questions which may have been troublesome to the repondent, but not obviously so to the interviewer, can be identified.

5. The Evaluation Plan

Much of the evaluation in an informal test is simply the use of common sense in reacting to problems identified by the feedback system. The lack of objective criteria for evaluating the questionnaire responses may be seen as a disadvantage of this type of testing. However, some quantification of the responses may be possible (e.g., tabulations of the number of "don't know," "refused," or "not applicable" responses to a question). These types of responses in addition to inconsistent and missing responses often indicate various questionnaire problems.

Simple frequency distributions may also be tabulated and compared to known distributions to help determine the appropriate response categories for a question. These tabulations can usually be performed clerically because of the small number of cases. If two different questionnaires have been used,

the data should be used for descriptive purposes only and cannot be used to make statistically significant comparisons between the questionnaires.

D. Operation

The evaluation of an informal test involving personal or telephone interviews can be hindered if steps are not taken to ensure that the questionnaire is administered properly. The persons conducting and observing interviews should understand the objectives of the test and the importance of not arbitrarily varying the questionnaire wording and administration. However, they should know how to probe by rewording questions or asking other questions when it is suspected that a response is inaccurate, inappropriate, or insufficient. Probing should only be used under circumstances approved by the questionnaire designer/researcher to provide further insight into potential questionnaire problems; when used, it should be noted as part of the feedback system.

The lines of communication between the questionnaire designers, observers, interviewers, and other project staff should be well-established to enhance the feedback. One major advantage of an informal test is the possibility of making on-the-spot revisions to the questionnaire as a result of the feedback. Because of the small number of people and questionnaires involved, any problems uncovered can be discussed at the end of one day's interviewing and changes made before the next day's interviewing begins. These changes and the rationale for making them should be carefully documented for use in evaluating the questionnaire's performance and for future use by others who are performing related work.

Following the data collection portion of an informal test, the information gathered through debriefings, observations, and tabulations of the survey data or evaluation forms should be examined to determine what changes should be made in the questionnaire. Thorough documentation of the process and any resulting questionnaire changes should be made for use by future researchers. Unfortunately, the test often only indicates that there is a problem; it does not provide the "correct" solution. For example, if a given question is not answered frequently in a test, there may be a problem with the wording. However, unless the interviewers or observers have probed to find out why the question is not being answered, the questionnaire designer might not have enough information to rephrase the question in a way which will elicit more responses.

E. Time Considerations

The amount of time required to conduct an informal test varies according to a number of factors. Assuming that the questionnaire has been drafted,[1] the total amount of time which should be allotted for the operational aspects of

---

[1]The time required to draft the questionnaire varies considerably depending on how much developmental work is necessary--for example, whether the survey has been conducted before or is totally new, or whether any of the developmental techniques described in Part II of this report have been used.

an informal test is approximately 3 to 4-1/2 months. This includes time for OMB approval[2] (during which manuals, training, and field procedures can be prepared, an interviewing site and a sample of respondents can be selected, and forms can be designed if necessary), selection and training of interviewers, reproduction of necessary materials, data collection, receipt of feedback through interviewer debriefing, completion of observer reports, etc., and summarizing the results. The variable factors which prohibit specification of an exact time frame include (1) the number of cases and interviewers, (2) the length of the interview and the distance between sample households, (3) whether materials can be duplicated in-house or must be sent to a printing company, (4) whether interviewer instructions, training materials, debriefing guides, and observer forms are written (the larger the number of sample cases, the more likely it is that these materials will be put in writing), and (5) whether materials have to be mailed to the interviewing site.

F. Cost Considerations

Relative to other types of field tests, informal tests are inexpensive data collection efforts. This, in addition to the relative speed with which they are conducted, contributes to their usefulness as tools for questionnaire design.

It is difficult to quantify a cost range for conducting an informal test.[3] However, the factors which contribute to the costs are (1) interviewing and field staff salaries (this is the major cost), (2) other professional salaries (i.e., questionnaire designers, observers), (3) travel and expenses for interviewers and observers (if the test is not being done locally), (4) forms design and/or reproduction of questionnaires, and (5) postage (if materials need to be mailed to the field).

G. Mode of Data Collection

Informal testing is an equally appropriate technique for use in the development of face-to-face and telephone questionnaires. The relationship between the mode of interviewing used in an informal test and that used in the final survey was discussed previously in section II, part B.

Since one of the positive features of informal tests is the opportunity for the interviewer and/or observer to converse with the respondent after the interview about problems which may have been encountered, this type of testing is not as useful in the development of a mail questionnaire where

---

[2]OMB approval is required for all data collection efforts that will involve more than nine respondents. OMB's role is to ensure that information collected is in the public interest, that respondent reporting burden is reasonable, and that certain statistical standards are met. OMB now (1983) requires 60 days to review requests for approval.

[3]A very tentative estimate of the cost range involved in conducting an informal test for a large-scale national survey is $5,000 to $30,000 (in 1983 dollars).

interviewers and/or observers are nonexistent. Formal testing may be more appropriate for a mail questionnaire, depending on time and cost constraints, because it should provide more useful results from statistical tabulations than would an informal test.

III. EXAMPLES

Two different types of informal tests were conducted to help develop the questionnaire for the 1980 National Survey of Fishing, Hunting, and Wildlife Associated Recreation (FHWAR). Although this survey had been conducted at 5-year intervals since 1955, it was acknowledged that the previous questionnaires contained some weaknesses. Specifically, there were needs for better data on "nonconsumptive users" of wildlife resources, such as birdwatchers, and on the economic value of hunting and fishing activities. As a result, informal testing was undertaken prior to the survey to develop techniques and questions in these areas (example 1) and to refine the questionnaire (example 2). The first example was chosen for this report because it shows how a series of tests may be used to make a decision on the best way to elicit the required information.

A. Example 1: 1980 National Survey of Fishing, Hunting, and Wildlife Associated Recreation

1. Objectives

Human Sciences Research Inc. (HSR) was selected by the survey's sponsor, the Fish and Wildlife Service (FWS) of the Department of the Interior, to perform the developmental work on collecting data to produce measures of the economic value of hunting and fishing.[4] Preliminary studies showed that existing valuation methodologies could be adapted for use in the survey. These methodologies required data on the location of hunting and fishing activities and on respondents' willingness to pay for participation in these activities. Therefore, the major objectives of HSR's work were to determine (1) the best method of asking questions to locate the site(s) used for hunting and fishing and (2) the best technique for getting respondents to put a dollar value on these activities. To achieve these objectives, HSR conducted a series of informal tests to compare two methods for determining hunting and fishing locations and to evaluate the use of a bidding game technique to obtain willingness-to-pay (WTP) data.[5]

---

[4]Selected portions of the following material have been excerpted from a report by Human Sciences Research Inc. (1980).

[5]It should be noted here that the use of a bidding game technique has limitations which discourage its use in some circumstances. One of the purposes of this informal test was to examine its feasibility in this circumstance. The description of the evaluation of the bidding game questions contained here is intended as an example of how questionnaire revisions are accomplished in an informal test, not as encouragement for others to use the bidding game technique.

To determine locations where fishing and hunting had occurred, the following questions were developed to contrast general and specific approaches:

General: Can you show me on this map where you hunted (fished)? Just tell me the number of the region outlined on this map.

Specific: Can you tell me the name of the lake, stream or nearest area where you hunted (fished)?

To get respondents to put a dollar value on hunting and fishing, a bidding game technique was used to determine willingness to pay for various activities. The game was played by using a series of questions to establish the actual cost of an activity and then determine the maximum amount the respondent would be willing to pay for it. For example:

1. What did a hunting license cost you in 1979? $ X

2. Would you continue to hunt if the license cost $ 2X ? (2X equals two times the actual cost provided in response to question 1.)

OR

1. About how much do you figure your total costs were in 1979? $ X

2. If your costs increased to $ 2X would you still go?

If the answer to question 2 was "No," the bidding game was stopped; if the answer was "Yes," the question was asked again inserting an amount that was three times higher than the actual cost. This question was repeated using an increased amount each time until a "No" answer was received. Five different question series, including the ones above, were developed prior to testing and three additional variations were developed during testing.

To aid in the development of the location and bidding game questions for the tests, a qualitative group interview session was held with several types of hunters and fishers. This session assured the researchers that respondents could describe or identify the places where they hunted or fished and could understand the purpose and technique of the bidding game. (See Chapter 3 for further explanation of qualitative group interviews.)

2. Technical and Operational Considerations

Four sequential rounds of testing were planned to allow refinements suggested in one round to be tested in the next one. It also permitted testing the questions in four regions of the country which have different hunting and fishing activities. Five questionnaires were designed for the first test, each containing only one version of the original WTP questions. For example, in one questionnaire, a bidding game about the value of license costs was played for each activity in which the respondent participated. This resulted in considerable redundancy if the respondent participated in several activities such as trout fishing, deer hunting, and duck hunting. It was feared that this might reduce the respondent's willingness to give a reliable response each time the bidding game was played. However, the idea of using

more than one version of the WTP questions in a single questionnaire (e.g., license costs for trout fishing and total trip costs for deer hunting) was tried in the second test and rejected as a feasible alternative.

As the tests progressed, less successful WTP questions were eliminated, the most effective were repeated, and new WTP questions were developed and tried. By the last test, the number of WTP questions had been sufficiently narrowed to make the administration and observation easier. Also, by the third test, the specific question used to identify hunting and fishing locations was deleted because of several disadvantages that were observed. This ability to modify the test questionnaire and procedures and subsequently test these modifications contributed greatly to the final questionnaire design.

To ensure that the samples contained known hunters and fishers, the sponsor (FWS) obtained access to the records of the fish and wildlife agencies in four States: Florida, Missouri, Maryland, and Washington. A sample of persons who had purchased hunting or fishing permits in the previous year was selected from the records. To further narrow the sample to the population of interest in the survey, a telephone screening process was used to determine which sample persons actually had participated in hunting or fishing activities within the past year. This process identified about 25 to 35 persons for personal interviews in each of the four States. Since the survey also planned to use a screening process, these informal tests provided an opportunity to observe the proposed procedures in action.

During the interviewing, several feedback mechanisms were used to provide information on the performance of the methods and questions employed. First, survey teams were used consisting of one of the researchers, who conducted the interview, and an observer. Using a researcher as the interviewer guaranteed that the objectives and content of the test would be thoroughly understood. Since observers were also very familiar with data needs, it was possible for them to conduct a brief post-interview discussion with a respondent when it seemed necessary to clarify a question or obtain more information. Also, respondents were encouraged to ask questions and to give their opinions following the interview.

Daily debriefings of the interviewers and observers were held, too, and led to changed procedures which were implemented the next day. Finally, after each test, meetings were held to discuss modifications to the questionnaire prior to the next test.

3. Results

Following the four tests, the contractor prepared a brief report on the research which summarized the main conclusions. The cost of the work was approximately $30,000[6] and resulted in the following recommendations and observations which were used in developing the survey questionnaire: (1) The best information on the location of hunting and fishing sites was obtained when a map was used to display wildlife management regions within

[6] This is a relatively small amount compared to the estimated $6 million cost for the 1980 survey.

the State and surrounding States. (2) The bidding game technique appeared to be a feasible way of getting respondents to assign a value to their activities. The cost-per-day and cost-per-season WTP questions seemed to work best in the bidding games. Since daily costs were often used to establish seasonal costs for an activity, it was decided that willingness to pay could be determined by using only the cost-per-day bidding game. The other bidding game questions did not work, because among other reasons, they were too abstract, required too many calculations or created suspicion of the interviewer's motives. (3) Responses to the bidding games could be substantially biased by the interviewer. An attempt should be made to minimize the interviewer effect by using a verbatim guide for the bidding game, developing standardized procedures, and providing thorough training on the purpose and technique. Also, the interviewers need to be informed about kinds of local wildlife and hunting/fishing regulations so they can conduct the interviews smoothly and avoid mistakes.

B. Example 2: 1980 National Survey of Fishing, Hunting, and Wildlife Recreation

1. Objectives

An informal test was also conducted to refine other (nonbidding game) questions for hunters and fishers and to assess the clarity and comprehensiveness of the proposed nonconsumptive user questions for the 1980 FHWAR. Whereas many of the questions for fishers and hunters had been used in the previous surveys, the questions for wildlife photographers, birdwatchers, and other observers of wildlife were relatively untested. The Bureau of the Census undertook this phase of the informal testing in preparation for conducting the 1980 survey. This example was selected because several different types of questionnaire problems were detected during the test. (See the results section for a description of the problems.)

2. Technical and Operational Considerations

The Bureau's test was designed to use the basic methodology selected for the survey, namely, a telephone screening interview with a household respondent followed by a detailed personal interview with each household member who was identified as a hunter, fisher, or nonconsumptive user. Three questionnaires were used in this process: (1) a screening questionnaire to identify persons for further questioning, (2) a detailed questionnaire for hunters and/or fishers, and (3) a detailed questionnaire for nonconsumptive users. Persons who were both hunters/fishers and nonconsumptive users were administered both detailed questionnaires.

The methodology for the test varied from the survey in that a judgmental (nonprobability) sample was selected to provide a sufficient number of participants for personal interviews. (The survey used a probability sample.) The sample was selected from a list of respondents who had been in a survey conducted by the Michigan State Department of Natural Resources in 1979 and were licensed to hunt or fish at that time. It was assumed that it would be impossible to reach many of these persons by telephone (wrong number, no answer, etc.) and that some of those reached would not be identified as hunters, fishers, or nonconsumptive users. Also, of those identified, some

would be unavailable for a personal interview. Therefore, approximately 400 persons were initially selected from the list to ensure that at least 100 persons would be identified for a detailed interview. In addition, 25 households were selected from another survey conducted by the Bureau to include some households where the presence of fishers and hunters was unknown.

Ten experienced Census Bureau interviewers were selected to enable the test to be completed within 5 days. A self-study guide was sent to the interviewers to familiarize them with the concepts and procedures which would be used in the test. Then, classroom training was held to discuss the test procedures and provide practice in administering the questionnaires in mock interview situations. In addition, the Bureau prepared a reference manual to assist the interviewers in administering the questionnaire.

To aid in the test evaluation, Bureau and FWS staff members accompanied the interviewers to observe and report on the detailed interviews. In addition, the interviewers were encouraged to report any problems in a debriefing session following the interview period. The questionnaire data were not processed; however, some clerical tallies were made for evaluation purposes.

3. Results

The cost of this test was approximately $20,000, and it took a little over 3 months to plan, conduct, and evaluate. The test results were issued as internal memorandums only and showed that the screening interview was successful in identifying hunters and/or fishers and nonconsumptive users who were eligible for the detailed interview. However, because of time constraints, the interviewers were unable to obtain detailed personal interviews from all the people identified by the screener. With a longer interview period, many more interviews could have been scheduled.

There were two major findings, based on subjective evaluations, regarding the screening questionnaire. First, it was observed that length was affecting cooperation. In the test, 10 out of 100 respondents refused to allow a personal visit interview because of the time it had taken to complete the screening questionnaire. Therefore, it was recommended that the screening questionnaire be shortened by dropping several questions which were unnecessary for screening purposes.

The second major finding was that although household respondents seemed to be able to identify hunters and fishers, they had more trouble identifying nonconsumptive users. It was thought that the loose definition of nonconsumptive users might be the cause; therefore, it was recommended that those screener questions be clarified.

The observers and interviewers detected several problems with the detailed questionnaires used in the personal interviews. In general, the questions seemed repetitious and wordy. To help the flow of the interview, changes in the interviewing techniques, skip patterns, and questionnaire format were suggested. Some problems with specific questions included (1) confusing wording, (2) deficient visual aids, (3) vague terms and concepts, and (4) missing answer categories. Appropriate improvements were suggested where possible. Clerical tallies of item nonresponses were also used to identify

problems with specific questions, and efforts were made to change the questions to elicit more answers. Also, it was felt that better interviewer training would have reduced the number of nonresponses in some of these cases.

Overall, it was noted that the structure of the detailed questionnaires led to potential double reporting of information; e.g., three reports of one trip which involved hunting, fishing, and nonconsumptive activities or three reports of the same trip by three family members who went as a group. On the other hand, trips originating from a vacation home were probably missed because of the wording of the introduction to this set of questions. This resulted in some suggestions for restructuring the questionnaire and rewording the introduction.

The revised questionnaire was used in the survey which was completed in 1981. The Fish and Wildlife Service used the results to prepare a national report and individual State reports for the 50 States. The national report was released in November 1982 and the primary users, namely fish and wildlife planners at all levels of government, have found the data generally accurate and useful. These favorable results were probably due, in part, to questionnaire improvements arising from the informal testing.

Chapter 6
Formal Testing

The methodologies described in previous chapters have relied primarily on subjective assessments of questionnaire design, describing the limitations and benefits of such approaches. To achieve more confidence in the final questionnaire, however, survey planners may undertake more formal testing, that is, field testing which depends on probability sampling for respondent selection and for which results can therefore be evaluated on a rigorous statistical basis.

Formal testing includes testing of various types, and two such variations are described in this chapter. One type, referred to here as a pilot study, is a prototype of the survey conducted to observe all of the proposed survey operations working together, including questionnaire administration--i.e., a dress rehearsal of the actual survey. It calls for developing a design which duplicates the final proposed survey design on a small scale from beginning to end, including plans for data processing and analysis.

The second type, a split sample test, is conducted specifically to determine the "best" of two or more apparently feasible alternative versions of the questionnaire or almost any aspect of survey operations. (The terms "split ballot" and "split panel" have also been used to refer to such tests; however, split sample more directly describes the actual design and avoids confusion over more common uses of the terms "ballot" and "panel.")

Distinctions among types of formal testing are not always clear. For example, some split sample tests do more in the way of data processing and data analysis "dry runs" than others, depending on time and budget constraints. Pilot studies may also incorporate minor tests of alternatives in either the questionnaire or in various survey procedures, as opposed to being conducted primarily to test those alternatives.

Despite the potential for overlap between the various formal testing techniques, a general description of each basic type is presented here.

I. PILOT STUDIES

A. Introduction

The pilot study method calls for developing a design which duplicates the final proposed survey design on a small scale from beginning to end. For

Pilot Studies Principal Contributor: Gemma M. Furno
Split Sample Testing Principal Contributor: Janice Olson

example, to the extent possible, the pilot study questionnaire is identical to the final one in content, wording, layout/print style, sequencing, etc., and the interviewing method matches that chosen for the final survey. However, some aspects may have to be different--conducting a nationwide pilot study for a planned national survey is often not practical, especially when the interview mode is a personal interview. In such cases, the pilot study may be limited to carefully chosen regions or cities. An alternative to consider would be to tie the pilot study to an already existing national survey, thus achieving national coverage for the pilot. This alternative must be carefully planned since the presence of the other survey may make it difficult for the pilot study to duplicate the final survey procedures or to clearly determine such factors as the cause of refusals.

The advantage of being able to discover and correct any errors or problem areas before the actual survey begins is self-evident. The increasing cost of data collection and concern with undue respondent burden makes it even more crucial that the final survey effort be as successful as possible. A less widely recognized benefit of a pilot study is the potential for minimizing the delay between the final survey and the availability of the results because the post-collection survey procedures and analysis plan have already been developed and tested in the pilot. The disadvantages are primarily ones of time and cost. There is often little time scheduled between the formulation of the survey plan and the final data collection effort. A pilot study is inherently more costly than earlier forms of testing because it encompasses all survey procedures, not only the questionnaire, and because the sample size is larger. In evaluating the trade-offs between advantage and disadvantage, it is the large, complicated, or repetitive survey which usually warrants the pilot study investment, i.e., situations where the efficient operation and meshing of each phase is especially crucial to the success and cost of the project.

The basic questionnaire design issues which are amenable to detection in a pilot study are varied. The use of this technique provides the opportunity to see how well the questionnaire performs in conjunction with other phases of the survey. For example, the data processing phase may reveal keying problems with the format or typographical errors in the precoded item numbers and/or answer categories. Usually, minor corrections or modifications in the layout will correct these problems and improve the efficiency and accuracy of the coding/keypunching. This type of accommodation to processing needs would not be possible if the data processing procedures, including computer programs, were left until after the final version of the questionnaire was printed or the data collected. Interviewer training can also reveal such problems as typographical errors in the questionnaire, awkward question wording, or concepts which need further clarification to ensure that interviewer and respondent error are minimized. Carrying the pilot study through the analysis phase serves as a final check that the questionnaire will provide the data in the form needed. For example, a modification is in order if the analysis calls for presenting age as a mean but the questionnaire collects the data in range categories rather than in exact ages. In summary, many of the errors detected in a pilot study could be found at an earlier stage, but often, for a variety of reasons, are not. In addition, there are some problems which only surface when the interrelationships between survey phases are field tested.

In addition to the "dry run" function, a pilot study can also provide a vehicle to perform minor tests of alternatives. In regard to the questionnaire, this does not mean that the alternatives have not been previously tested, but rather, that more data are needed before a final decision is made. A pilot study is usually not the time to try out new questions or approaches. The results represent the final chance to fine tune the questionnaire. If previous testing has been done, it is highly unlikely that the pilot study will result in major changes to the questionnaire. However, if it does, further testing is indicated before the actual survey is conducted.

B. Method

1. Personnel and Skill Requirements

The magnitude of a pilot study and the use of a probability sample necessitate the involvement of larger numbers of personnel with a wider variety of skills than has been the case in any of the techniques described thus far. In addition to a project manager who performs the operation coordination function (described in Chapter 5 on informal testing), the services of a sampling statistician are required to ensure that proper sample selection procedures are employed. Expertise in the planning and execution of data processing tasks (i.e., editing, coding, keying) and skill in data analysis are also called for, although not necessarily in the same people.

To accommodate the large number of interviews involved in a pilot study, more interviewers are also necessary (for an interviewer-administered survey). Unlike previous stages of testing in which researchers/questionnaire designers can be used to conduct the interviews, the interviewers in a pilot study should be the same as those planned for the final survey. Depending on the geographic dispersion of the sample, one or more interviewing supervisors might be considered to keep track of the interviewing workload, to receive feedback from editors about the quality of interviewers' work, and to keep the interviewers informed about the quality of their work.

2. Selection of Respondents

When selecting the sample, the possibility of the same respondents falling into both the pilot study sample and the final survey sample should be avoided, because of conditioning effects and respondent burden. (Obviously in a census this is not possible.) Overlap can be avoided if both samples are selected at the same time, provided one does not expect the frame to become out of date between the time of the pilot and the final survey. For example, the sample for the final survey could be selected first and then a second sample drawn from the remaining elements of the universe, or one sample large enough for both operations could be drawn with subsequent subsampling for the pilot study. The end result is the testing of the sample design by actually drawing at least one sample. The sample size must be large enough to support the pilot study evaluation plan (item frequency counts, results of different alternatives tested, etc.) and to adequately test the primary analysis plan for the final data.

3. Preparation

In addition to the preparatory work involved in choosing interviewing sites and selecting a sample, interviewers working on the pilot study have to be selected and trained.

The selection and training of survey personnel (supervisors, clerks, interviewers, etc.) should follow the same procedures planned for the final survey. This is especially important if prior testing utilized only experienced interviewers or if the researchers served in this capacity. Often the pilot survey personnel are the same people or a subset of those who will work on the final survey. The pilot study can help identify any major recruitment problems or training deficiencies. If a pilot study for a nationwide survey is conducted in a city or region, one would have to consider whether recruitment in that area could be considered "typical."

Before interviewing can begin, sufficient copies of the questionnaire must be available for all the sample cases. Depending on the sample size and the time available, these can be either xerographic reproduced or printed copies.

4. Operation

Once the interviewers have been recruited and trained, the interviewing phase of the pilot study proceeds the same as for any other data collection effort involving a verbatim questionnaire. In addition, observers may accompany the interviewers in the field to gain first-hand knowledge of how well the questionnaire works. At the end of the survey period, evaluation forms may be completed or debriefing sessions can be arranged with the interviewers and/or observers to get feedback about their perceptions of the interviews. (See Chapters 8 and 9 for descriptions of the objectives and procedures involved in observation and interviewer debriefing.)

Researchers sometimes have a tendency to concentrate their energies on the questionnaire and collection phase of a survey to the neglect of the data processing operation. The pilot study concept, by demanding that a miniature survey be done from beginning to end, counteracts this inclination. As mentioned previously, the pilot study can show how the questionnaire should be modified to accommodate processing needs. A successful pilot study confirms that all the interrelated steps in the data processing phase (e.g., check-in, initial editing, coding, keypunching, transfer to tape, computer edits) are being coordinated in the most efficient manner. Bottlenecks are revealed and corrected while additional time-saving features may be discovered.

The final stage of any survey is the analysis of the data. To carry a pilot study to its logical conclusion, there should be an analysis of the data collected to the extent that the basic analytical design for the final survey data is tested. This allows researchers their first opportunity to see if the prior survey phases have produced data which are compatible with the analytical design and if that design is realistic (i.e., whether the survey will yield the type of data and in the expected distributions for any models, or other such analytic tools, to work properly). Adverse results from the data could suggest adjustments to the questionnaire, to various survey

procedures such as the edits or imputation criteria, or to the analytical design itself.

It should be acknowledged that it is at this final stage of data analysis that a pilot study usually falls short, more often because the analysis is not done, rather than because it reveals any defects. Efforts are more likely to be focused on evaluating the pilot study itself than on duplicating the analysis plan of the final survey, which unfortunately may not even be completed by this stage. However, the further a pilot study is pushed, the greater the potential for discovering and correcting individual errors and deficient choice of both survey and analytical methods before the final survey is conducted.

The evaluation of a pilot study requires careful planning as there is usually little time to identify problems and correct them. To avoid being overwhelmed by data, one should decide in advance what key indicators of the field work and processing will be requested and what observational feedback system to employ. The following list suggests areas most closely related to the questionnaire.

- Observation of training and interviewing
- Debriefings of interviewers and observers
- Simple frequency counts of all answer categories
- Survey and item response rates
- Questionnaire edit failures (omissions/inconsistencies detected in the clerical or machine clean-up of the data)
- Interview time
- Field costs

Often the most obvious way to evaluate the interrelationship between the various phases is to see what problems surface as the pilot study proceeds.

5. Time Considerations

The time required for a pilot study is considerably longer than that involved in an informal test. Once a questionnaire is available (either drafted from scratch or revised based on the results of an informal test), approximately 5 to 10 months should be allocated for the pilot. This allows time for OMB approval[1] (during which manuals, training, and field procedures can be prepared, interviewing sites and samples can be selected, and forms can be designed), selection and training of interviewers, printing necessary materials, data collection, clerical editing, coding, keying, programming,

---

[1]See section II, part E in Chapter 5 for information on the OMB approval process.

processing and analyzing the data, receiving feedback from subjective assessment tools, and summarizing the results.

This time frame may be shortened somewhat to accommodate a tighter schedule if additional personnel are available to help program or tabulate the data and to examine the data and interviewer/observer reports for ideas regarding the design of the questionnaire.

6. Cost Considerations

It is difficult to quantify a cost range for a pilot study.[2] However, it may be helpful to list the factors which add to the total cost. These elements (presented in decreasing order) include (1) salaries--for interviewers, planning staff (supervisors, interviewer trainers, etc.), computer programmers, clerks, and data keyers; (2) travel and expense costs for interviewers, supervisors, observers, and any other travellers; (3) costs for forms design and printing or other reproduction; (4) costs for data processing (clerical operations, keying and verification, tape preparation, computer programming, and data analysis runs) and postage, when applicable.

7. Mode of Data Collection

Pilot studies are appropriate techniques for use in developing face-to-face or telephone interview instruments or mail questionnaires. In addition, the (relatively) small amount of labor involved in pilot studies for mail surveys makes them much less costly and increases their practicality compared with pilots for face-to-face and telephone surveys. Unlike informal tests which are less useful for mail surveys than for other modes of interviewing, the emphasis on statistical evaluation of the results of formal tests makes them particularly suited for use with mail surveys.

C. Example: 1977 National Survey of Crime Severity National Field Test

The National Survey of Crime Severity (NSCS) was sponsored by the Law Enforcement Assistance Administration as part of a grant project conducted by the University of Pennsylvania Center for Studies in Criminology and Criminal Law. The survey was administered by the Census Bureau as a supplement to the National Crime Survey (NCS). The survey results were used to create a crime-seriousness weighting system which resulted in the construction of a crime-seriousness scale. This scale allows policymakers and researchers to determine changes in the total severity of crime and to focus attention on crimes perceived as more serious than others.

The NSCS National Field Test was chosen as a pilot study example because the final survey (50,000 respondents) required a considerable commitment of resources and utilized a complicated technique which had not been fully tested before in personal interviews with a sample of the general population. It also illustrates the use of a pilot study to test final questionnaire alternatives: in this case, which scoring technique would prove easiest to

---

[2] A very tentative estimate of the cost involved in conducting a pilot study is $10,000 to $50,000 (in 1983 dollars).

administer and provide the best data for the construction of a seriousness scale or index. Except for interviewer training (2-hour, self-study vs. planned 1/2 day for final survey), the pilot duplicated all aspects of the final survey plan.

The major problems faced in developing the NSCS were creating a workable set of respondent and interviewer instructions to administer the task[3] to a sample of the general population, simplifying and neutralizing the wording to over 250 crime vignettes, and determining if any vignettes described behavior too sensitive or embarrassing to ask about in an interview.

Initially, the NSCS went through two stages of informal testing in November 1976 (4 rounds of local developmental work involving 25 interviews by research teams and a field test in 9 areas across the country involving 110 interviews conducted by experienced crime survey interviewers accompanied by research observers). This work culminated in a national field test in February 1977 in which 2,452 completed interviews were obtained. The final survey was conducted from July to December (1977) and yielded a little over 50,000 completed interviews.

Note that this time schedule was extremely tight and may not be typical of similar projects. To accomplish three distinct testing phases and to prepare for the final survey in only 8 months called for intense effort by many people, especially during the peak periods when quick evaluation of the test results was required.

The NSCS National Field Test was conducted as a supplement in the expiring sample rotation groups of the National Crime Survey. Since no more interviews were scheduled at these households, the possibility of interviewing the same people in the final survey was avoided. The plan to administer the final survey as a supplement to a national survey provided the added benefit of economically achieving national coverage for the pilot study.

As mentioned previously, there were originally over 250 items for which scores were to be obtained. By the time of the pilot, this list had been pared to about 200, but it was still far too large to ask each respondent to score. To keep respondent burden at a reasonable level, most of the items were randomly distributed among 12 different questionnaire versions. (See Figure 1 for a copy of questionnaire version 1.) However, a subset of 12 core items, essential to later severity scale construction, was included more frequently than the rest to obtain a larger number of cases. Each version contained a prescored reference item (bicycle theft at 10), 3 practice items intended to help respondents understand the procedure and the

---

[3]Each respondent was read a set of brief descriptions of various types of criminal acts. For each description, or vignette, the respondent assigned a numerical score to its relative seriousness in comparison to a prescored reference event (a bicycle theft scored at 10). The variables covered by the vignettes included type of crime, amount of loss or damage, extent of injury, presence and type of weapon, type of victim (private person, commercial, public), use of force or intimidation, and type of offender (juvenile/adult, commercial, public).

Figure 1. Questionnaire Version 01 in Pilot Study

O.M.B. No. 43-S76010; Approval Expires April 30, 1977

FORM NCS-201(X)
(1-4-76)

U.S. DEPARTMENT OF COMMERCE
BUREAU OF THE CENSUS
ACTING AS COLLECTING AGENT FOR THE
LAW ENFORCEMENT ASSISTANCE ADMINISTRATION
U.S. DEPARTMENT OF JUSTICE

NATIONAL SURVEY OF CRIME SEVERITY

VERSION 1

NATIONAL CRIME SURVEY SUPPLEMENT

NOTICE – Your report to the Census Bureau is confidential by law (U.S. Code 42, section 3771). All identifiable information will be used only by persons engaged in and for the purposes of the survey, and may not be disclosed or released to others for any purpose.

Sample	Control number: PSU	Segment	CK	Serial	Household number	Version number
JO___						01

Respondent
Line No. | Name

VERSION 01

A. Interviewer identification
Code | Name

B. Interviewer introduction used
1 ☐ Introduction 1
2 ☐ Introduction 2
Use only one introduction per household. Alternate introductions between households.
9 ☐ OFFICE USE ONLY

C. Type of interview
1 ☐ Personal
2 ☐ Telephone
9 ☐ OFFICE USE ONLY

D. Length of interview
Time began a.m. p.m. | Time ended a.m. p.m. | No. of minutes

E. Date completed

Interviewer Check Item
1 ☐ Understood task
2 ☐ Did not understand task
3 ☐ Not sure; don't know
Explain on reverse
9 ☐ OFFICE USE ONLY

F. REASON FOR NONINTERVIEW

Interviewer Instructions: *Complete one supplement form if entire household is a noninterview on National Survey of Crime Severity only. Otherwise, complete one supplement form for each household member 18+. (If Type Z Noninterview fill that section only.)*

NSCS Type A Noninterview
1 ☐ Noninterview Supplement Only
Reason: _____

Type Z Person
Reason:
2 ☐ Type Z Noninterview on NCS
3 ☐ Proxy Interview on NCS
4 ☐ Refused NSCS (Supplement Only)
5 ☐ Other – *Specify*
9 ☐ OFFICE USE ONLY

INTERVIEWER INSTRUCTIONS ▶ Interview all household members 18 years and over (proxy interview not acceptable)

CONTINUATION OF INTERVIEWER INTRODUCTION

To give you some practice in how to do this, consider the following situation: "An offender robs a person. The victim is injured but not hospitalized." What number would you give to this situation to show how serious you think it is compared to the bicycle theft with a score of 10? *(Obtain answer)*

The next situation is: "A juvenile plays hookey from school." Compared to the bicycle theft with a score of 10, how serious do you think this is? *(Obtain answer)*

The next situation is: "An offender stabs a person to death." Compared to the bicycle theft with a score of 10, how serious do you think this is? *(Obtain answer)*

1. An offender steals a bicycle parked on the street	10
2. An offender robs a person. The victim is injured but not hospitalized	
3. A juvenile plays hookey from school	
4. An offender stabs a person to death	

(Interviewer: *Review answers to see if respondent understands that less serious crimes should be given a number less than 10 while more serious crimes should be given a number greater than 10. If there is any doubt about respondent's comprehension of the task repeat the instructions as needed* ***before continuing.)***

The remaining situations, like the ones we have just done, vary widely from very minor to very serious. Score them in the same way by comparing each one to the bicycle theft before arriving at your answer. *(PAUSE)* If you do not consider something a crime give it a score of zero. If you think it's equally as serious give it a score of 10. *(PAUSE)* Try to think about the seriousness of the act itself without adding any other details.

Compared to the bicycle theft scored at 10, how serious is . . .	
5. An offender kidnaps a victim	
Compared to the bicycle theft scored at 10, how serious is . . .	
6. A large company illegally conspires with other companies to fix the retail prices of their products	
7. An offender steals property worth $10 from outside a building	
8. An offender robs a person of $1,000 at gunpoint. The victim is wounded and requires treatment by a doctor but not hospitalization	
9. An offender conceals the identity of others that he knows have committed crimes	
10. A company pays a bribe of $10,000 to a congressman so he will vote for legislation favoring the company	
Compared to the bicycle theft scored at 10, how serious is . . .	
11. An offender takes part in a dice game in an alley	
12. An offender intentionally injures a victim. As a result, the victim dies	
13. An offender walks into a public museum and steals a painting worth $1,000	
14. An offender injures a victim. The victim is treated by a doctor but is not hospitalized	
Compared to the bicycle theft scored at 10, how serious is . . .	
15. An offender does not have a weapon. He threatens to harm a victim unless the victim gives him money. The victim gives him $10 and is not harmed	
16. An offender smokes marijuana	
17. An offender breaks into a display case in a jewelry store and steals $1,000 worth of merchandise	
18. An offender tries to entice a minor into his car for immoral purposes	
Compared to the bicycle theft scored at 10, how serious is . . .	
19. An offender, using force, takes $10 from a victim. The victim is hurt and requires hospitalization	
20. An offender sets fire to a building causing $100,000 worth of damage	
21. A factory knowlingly disposes of its harmful waste in a way that pollutes the water supply of a city. As a result, 20 people die	
22. An employer orders one of his employees to commit a crime	
Compared to the bicycle theft scored at 10, how serious is . . .	
23. A man beats his wife with his fists. She requires hospitalization	

19 or 20 items. In general, the order of all items on each version was randomized (except the bicycle theft and the practice items). In a few cases, the random order had to be adjusted to eliminate chance clusterings of similar types of offenses.

Two introductions using different scoring techniques were tested in the pilot. These two approaches represented the best alternatives to emerge from the informal testing. Introduction 1 used an open-ended scaling technique known as magnitude estimation to obtain ratios (i.e., judge how much more or less serious a vignette is than the prescored bicycle theft) while Introduction 2 asked respondents to perform essentially the same task, but within a more familiar approach of a 0-1,000 range. (See Figure 2 for the wording of the introductions.) Interviewers were instructed to alternate the introductions between households.

There were four general objectives in the NSCS National Field Test:

1. Scoring Technique. The sample was designed to produce enough cases to evaluate whether the scoring technique in Introduction 1 could be administered/comprehended successfully and whether the data collected would differ significantly from that obtained using the scoring technique in Introduction 2.

   This test of alternatives was necessitated by the inability of either approach to emerge as the obvious choice based on earlier testing. Most everyone agreed that the 0-1,000 range approach was simpler to administer and comprehend. But the proponents of the open-ended scale in Introduction 1 (the survey sponsors) pointed out that the construction of a seriousness scale or index depends on fitting the perceived seriousness of various crimes with a power function. The development of a power function requires that respondents reply in terms of ratios (e.g., how many times more or less serious a vignette is compared to the reference). The 0-1,000 range has the disadvantage of not asking directly for ratios and possibly restricting the variation of answers.

   The proponents of the approach used in Introduction 2 felt that under magnitude estimation many people still chose to use their own closed-ended scale, regardless of the instructions because responses rarely went over 1,000. While the prescored reference was helpful, it was unclear how many were actually using it to respond with ratios.

2. Basic Procedures and Computer Programs. The basic procedures were complicated enough to require a dress rehearsal to check that all 12 versions could be assigned and processed correctly and that the interviewers could administer the NSCS without biasing the results. The formal test provided enough cases to adequately test the computer programs.

3. The 12 Questionnaire Versions. Previous testing had not produced enough cases to fully test the wording, ordering, and possible item sensitivity on all 12 versions.

4. Analysis Plan. Before committing resources to collect data from over 50,000 respondents, the basic plan of analysis, i.e., the construction of

Figure 2. Introductions Tested in Pilot Study

FORM NCS-200(X)
(12-22-76)

U.S. DEPARTMENT OF COMMERCE
BUREAU OF THE CENSUS
ACTING AS COLLECTING AGENT FOR THE
LAW ENFORCEMENT ASSISTANCE ADMINISTRATION
U.S. DEPARTMENT OF JUSTICE

INTERVIEWER INTRODUCTIONS
NATIONAL SURVEY OF CRIME SEVERITY

NATIONAL CRIME SURVEY SUPPLEMENT

INTRODUCTION (1)

I would like to ask your opinion about how serious YOU think certain crimes are compared to others. *(PAUSE)* I will read a list of situations that may be considered crimes. When I read each situation give me a number that indicates how serious YOU think each one is. There are no right or wrong answers. *(PAUSE)*

The first situation, "An offender steals a bicycle parked on the street," has been given a score of 10 to indicate its seriousness. *(PAUSE)* Use this first situation as a standard by which to judge all the other situations. For example, if you think a situation is 20 TIMES MORE serious than the bicycle theft, the number you tell me should be around 200 *(PAUSE)* or if you think it is HALF AS SERIOUS, the number you tell me should be around 5 and so on. *(PAUSE)* You may use ANY numbers from 0 on, as high as you want to go. *(PAUSE)* Consider the seriousness of each situation AS STATED without adding ANY OTHER details. Try to think about the seriousness of the ACT ITSELF.

INTRODUCTION (2)

I would like to ask your opinion about how serious YOU think certain crimes are compared to others. *(PAUSE)* I will read a list of situations that may be considered crimes. When I read each situation give me a number from 0 to 1000 that indicates how serious YOU think each one is. There are no right or wrong answers. *(PAUSE)*

The first situation, "An offender steals a bicycle parked on the street," has been given a score of 10 to indicate its seriousness. *(PAUSE)* As I read the other situations, give me any number from 0 to 1000 but compare them to the bicycle theft with a score of 10 before arriving at your answers. *(PAUSE)* For example, numbers less than 10 should be used for situations you consider less serious than stealing a bicycle, *(PAUSE)* while numbers greater than 10 should be used for situations you consider more serious. *(PAUSE)* Consider the seriousness of each situation AS STATED without adding ANY OTHER details. Try to think about the seriousness of the ACT ITSELF.

a severity index based on responses from the general population, needed to be confirmed. Previous indexes were based on data collected in small classroom experiments, with written rather than oral administration, and with respondent groups who were not representative of the general population.

Many tools were used to evaluate the pilot study. In addition to the actual severity scores, other information such as type of interview (telephone vs. personal visit), length of interview, noninterview reason, interviewer's opinion of respondent's comprehension, and introduction used were collected for each sample case. These data were used to produce simple computer frequency counts and cross-tabulations. Observation/evaluation reports were completed by interviewers and observers and debriefings of both groups were held. Each major tool used to evaluate the pilot study is described below.

1. Analysis of Severity Scores. The University of Pennsylvania (one of the sponsors) used the pilot study data to test whether the NSCS could indeed generate a scale and the form it would take. This exercise not only confirmed the basic analysis plan, but provided the criteria to judge which scoring technique provided the most valid input.

2. Simple Frequency Counts, Statistics, Cross-Tabulations. In general, the goal was to investigate whether any differences in such things as length of interview, noninterview rate, range of scores used, and number of different scores used appeared to be a function of the scoring technique. In addition, hand tallies were made of the noninterview reasons and the interviewers' written notes on questionnaires which indicated that some respondents may not have understood the task. The mean score for each item was used to construct bar-graphs for each version to check that the mix of items and their order on a version had not produced an anomaly (i.e., a version with too many less serious crimes or crimes of about the same severity or type listed together).

3. Observer Debriefing and Observation Reports. Eight staff members/researchers observed 65 NSCS pilot interviews in Boston, Chicago, Detroit, New York, Miami, Philadelphia, Washington, D.C., and Trenton. A debriefing meeting was held and the observation reports were hand tallied separately for each introduction (i.e., scoring technique).

4. Interviewer Debriefing and Evaluation Forms. Taped interviewer debriefings were held in Detroit, Chicago, and New York. All interviewers were requested to fill an NSCS evaluation form after completing their interviewing assignment. Topics covered included which scoring technique was preferred, item sensitivity, respondent comprehension, and wording problems on specific vignettes.

The major result of the pilot study was the choice of the scoring technique used in Introduction 1 (magnitude estimation) for the final survey. Analysis of all the data (severity scores, constructing the index, observations, simple frequency counts, etc.) showed that the magnitude estimation approach was slightly more difficult to administer and comprehend than the 0-1,000 scale. However, the closed-ended scale in Introduction 2 suffered from a tendency to cluster scores at the upper range limit of 1,000, thus artificially

compressing the ratios of offenses perceived to be extremely serious while causing less seriously perceived offenses to be overvalued. Both scales generated by the pilot study data could be fit by a power function (the premise of the analysis plan) but that fit was marginally better for magnitude estimation as compared to the 1,000 point scale. This advantage outweighed any administration and comprehension difficulties.

There were other results as well. Figure 3 is a copy of Version 1 used in the final NSCS survey. A comparison with the version used in the pilot (Figure 1) reveals the following changes to the questionnaire:

1. The entire introduction was shortened.

2. The word "practice" was not used in asking respondents to score the three preliminary items intended to help them rehearse the procedure and ensure their comprehension of it. Several respondents felt they did not need practice, and this put the interviewer in an awkward position.

3. The word "offender" was removed from the vignettes, and other words such as "knowingly," "illegally," and "unlawfully" were used to clarify the intent of the item. There was a feeling that the word "offender" confused some respondents and biased others to give higher scores. The vignettes were fine tuned further by simplifying a word or repositioning an occasional item.

4. Categories of noninterview reasons were developed from the hand tally of write-in entries on the pilot.

Other procedural changes were also made:

1. The test indicated that a language problem existed for Spanish-speaking respondents. As a result, Spanish versions of all 12 questionnaires were prepared. Unfortunately, there was not time to test them. The translations were reviewed by several people familiar with different Spanish idioms.

2. The keying instructions allowed a maximum of six digits for each score. Any score of 1 million or more was clerically assigned a score of 999,999. Examination of the cases in which extreme values were reported revealed situations where the coder mistakenly assigned something less than a six-digit string of nines to scores of a million or more. This resulted in the item receiving a lower numerical score than the maximum 999,999. The final survey plan included an edit to ensure that this did not happen.

3. The majority of the observers felt that, in general, the interviewers did a good job and knew the NSCS procedures. However, the experience of two observers suggested that the pilot instructions (2-hour, self-study) were inadequate and that the half-day classroom training for the final should stress developing the interviewing skills needed to administer the NSCS in a correct, nonbiasing manner.

Figure 3. Questionnaire Version 01 in Final Survey

Form Approved: O.M.B. No. 43-S76010

FORM NCS-201
(4-14-77)

U.S. DEPARTMENT OF COMMERCE
BUREAU OF THE CENSUS
ACTING AS COLLECTING AGENT FOR THE
LAW ENFORCEMENT ASSISTANCE ADMINISTRATION
U.S. DEPARTMENT OF JUSTICE

NATIONAL SURVEY OF CRIME SEVERITY
VERSION 01
NATIONAL CRIME SURVEY SUPPLEMENT

NOTICE – Your report to the Census Bureau is confidential by law (U.S. Code 42, section 3771). All identifiable information will be used only by persons engaged in and for the purposes of the survey, and may not be disclosed or released to others for any purpose.

A. Sample (cc 4)	B. Control number (cc 5) PSU	Segment	CK	Serial	C. H.H. No. (cc 2)	D. Version No.
JO___						01

E. Respondent Line No. | Name

F. Interviewer identification Code | Name

G. Date completed

H. Type of interview
1 ☐ Personal
2 ☐ Telephone
3 ☐ Not applicable
9 ☐ OFFICE USE ONLY

I. Was anyone else present during interview?
1 ☐ Yes – All
2 ☐ Yes – Part
3 ☐ No
4 ☐ Not applicable
9 ☐ OFFICE USE ONLY

J. Reason for noninterview
1 ☐ Type Z noninterview on NCS
2 ☐ Proxy interview on NCS
3 ☐ Refused NSCS (supplement only)
4 ☐ Language difficulty
5 ☐ Could not understand instructions –*Explain on reverse side*
6 ☐ Other – *Specify*

OFFICE USE ONLY	K.	L.	M.	N.	O.	P.

INTERVIEWER INSTRUCTION ▶ *Interview all household members 18 years and over (proxy interview not acceptable)*

INTRODUCTION – I would like to ask your opinion about how serious YOU think certain crimes are.

The first situation is, "A person steals a bicycle parked on the street." This has been given a score of 10 to show its seriousness. *(PAUSE)* Use this first situation to judge all the others. For example, if you think a situation is 20 TIMES MORE serious than the bicycle theft, the number you tell me should be around 200 *(PAUSE)* or if you think it is HALF AS SERIOUS, the number you tell me should be around 5 and so on. *(PAUSE)* There is no upper limit; use ANY number so long as it shows how serious YOU think the situation is. *(PAUSE)* If YOU think something should not be a crime, give it a zero. *(PAUSE)*

Consider the following situation: "A person robs a victim. The victim is injured but not hospitalized." What number would you give to this situation to show how serious YOU think it is compared to the bicycle theft with a score of 10? *(Obtain answer)* . . . . . .

"A person under 16 years old plays hooky from school." Compared to the bicycle theft with a score of 10, how serious do YOU think this is? *(Obtain answer)* . . . . . .

"A person stabs a victim to death." Compared to the bicycle theft with a score of 10, how serious do YOU think this is? *(Obtain answer)* . . . . . .

1. A person steals a bicycle parked on the street	10
2. A person robs a victim. The victim is injured but not hospitalized	
3. A person under 16 years old plays hooky from school	
4. A person stabs a victim to death	

Let's go over these first few answers to be sure I have recorded them correctly. You feel that a robbery in which the victim is injured is (more/less/as) serious (than/as) the bicycle theft, *(PAUSE)* and that playing hooky is (more/less/as) serious (than/as) the bicycle theft; is that correct? *(PAUSE)*

INTERVIEWER INSTRUCTION: *Stop and resolve any misunderstandings about the instructions. Make any changes to the practice scores as needed.*

Score the remaining situations in the same way by comparing each one to the bicycle theft. There are no right or wrong answers. Remember, you may use any numbers, as high or low as you wish. *(PAUSE)*

COMPARED TO THE BICYCLE THEFT SCORED AT 10, HOW SERIOUS IS . . .	
5. A person kidnaps a victim.	
COMPARED TO THE BICYCLE THEFT SCORED AT 10, HOW SERIOUS IS . . .	
6. Several large companies illegally fix the retail prices of their products.	
7. A person steals property worth $10 from outside a building.	
8. A person robs a victim of $1,000 at gunpoint. The victim is wounded and requires treatment by a doctor but not hospitalization.	
9. A person conceals the identity of others that he knows have committed a serious crime.	
10. A company pays a bribe of $10,000 to a legislator to vote for a law favoring the company.	
COMPARED TO THE BICYCLE THEFT SCORED AT 10, HOW SERIOUS IS . . .	
11. A person takes part in a dice game in an alley.	
12. A person intentionally injures a victim. As a result, the victim dies.	
13. A person walks into a public museum and steals a painting worth $1,000.	
14. A man forcibly rapes a woman. No other physical injury occurs.	
15. A person does not have a weapon. He threatens to harm a victim unless the victim gives him money. The victim gives him $10 and is not harmed.	

COMPARED TO THE BICYCLE THEFT SCORED AT 10, HOW SERIOUS IS . . .	
16. A person smokes marijuana.	
17. A person breaks into a display case in a store and steals $1,000 worth of merchandise.	
18. A person knowingly lies under oath during a trial.	
19. A person, using force, robs a victim of $10. The victim is hurt and requires hospitalization.	
20. A person intentionally sets fire to a building causing $100,000 worth of damage.	
COMPARED TO THE BICYCLE THEFT SCORED AT 10, HOW SERIOUS IS . . .	
21. A factory knowingly gets rid of its waste in a way that pollutes the water supply of a city. As a result, 20 people become ill but none requires medical treatment.	
22. An employer orders one of his employees to commit a serious crime.	
23. A person steals property worth $1,000 from outside a building.	
24. A man beats his wife with his fists. She requires hospitalization.	
25. A person plants a bomb in a public building. The bomb explodes and 20 people are killed.	

Q. To help us understand peoples' scores, I would like to ask an additional question. *(PAUSE)* **BEFORE** I gave you the last item to score, did you have an upper limit or a highest number in mind that you wouldn't go over?

1 ☐ No – End Interview
2 ☐ Yes – **What was it?**_____ *(Explain on reverse side any special circumstances, then end interview.)*

VERSION 01

In general, the pilot did an excellent job of revealing potential problems or weaknesses in the survey plan. However, there were two problems on the final survey which were not revealed in the pilot, one because the formal test did not follow the final survey plan and the other because of human error.

The first problem involved resolving mismatches in the final survey when the severity questionnaire was matched to the NCS questionnaire in order to pick up respondents' demographic and victimization characteristics. To save time, and because the demographic characteristics were not needed to construct a rough severity scale or to evaluate the scoring technique, a match was not made with the pilot study data. The only items used to perform the computer match in the final survey were the household and person identification numbers. This proved somewhat inadequate and the additional variables of age and/or sex on the NSCS questionnaire would have been helpful. If a match had been done for the pilot, the problem would have been discovered and the suggestion to add another match variable would have been implemented.

The second problem involved a miscalculation of the nonresponse rate on the pilot. The procedure used to count cases where the household was interviewed for NCS but all eligible respondents refused the NSCS led to an underestimate of the NSCS nonresponse rate. The procedure was correctly changed for the final survey without fully realizing how much the pilot nonresponse had been underestimated. A higher nonresponse rate than anticipated on the final survey was the result. In fact, the first 2 months (July and August) produced rates so high (17.1 percent and 15.8 percent) that the cases were sent back to the field for follow-up. After that, noninterview rates were monitored more closely than usual and offices were not permitted to close out until their nonresponse level was deemed acceptable. The average nonresponse rate for the final survey was 13 percent. If this problem had been recognized in the pilot, all the special procedures would have been in place and the training would have emphasized that a high rate was unacceptable.

II. SPLIT SAMPLE TESTING

A. Introduction

A split sample test is defined by Jabine and Rothwell (1970) as "a controlled experiment in which the treatments to be compared and analyzed consist of versions of a questionnaire which differ in one or more respects but have the same data objectives." The purpose is to determine the "best" alternative among two or more apparently feasible alternatives. Thus, the main feature which distinguishes split sample testing is the experimental design which is incorporated into the data collection process. In a simple split sample design, half of the sample cases might receive one experimental treatment and half, the other. In a test involving two experimental treatments, the framework might be a 2 x 2 factorial design with each of the two treatments in each experiment being tested on half of the sample.

The decision to undertake a split sample test may arise from a variety of sources. Designers may need greater confidence in (or more solid justification of) the viability of a proposed questionnaire and the quality of the data it would provide. Decisions to test alternative treatments may arise

from previous survey research results, informal tests, and questionnaire evaluation methods such as observation and interviewer debriefing (discussed in Chapters 8 and 9) which may give uncertain or conflicting assessments of the best approach. In addition, designers and/or researchers may be uncertain or may disagree about the best strategy for obtaining the most valid and reliable responses.

A split sample test can be designed to suggest optimal strategies for dealing with a wide variety of the questionnaire design issues outlined in Chapter 1. Such diverse aspects of questionnaire design as question wording, question sequencing, and procedural issues can be manipulated in experimental treatments. This allows investigation of the effects of such things as question length, question context, questionnaire flow, location of sensitive items, choice of respondent rules, mode of interviewing, and length of reference period.

For this reason, the technique has been used heavily in methodological studies designed primarily to advance basic knowledge of questionnaire design and the survey research process. For example, the work reported by Schuman and Presser (1981) on the nature of attitude questions relied on several hundred experiments in more than 30 surveys, mostly "piggybacking" on telephone surveys conducted by the University of Michigan's Survey Research Center.[4] Split samples also played a major role in research on response effects for threatening and nonthreatening questions described in Bradburn and Sudman (1979).

Because of interest in assessing trends over time and in the comparability of results across surveys, split sample approaches may have an important transitional objective in repetitive or recurrent surveys. In such surveys, a key concern is that <u>any</u> change in the questionnaire or procedures may have unknown effects on <u>other</u> data items, not just the items being added or refined. When that concern is balanced against the need for new information or against known problems with traditional items or approaches, the result is often a split sample approach in which a random portion of the respondents might receive the "old" questionnaire and the rest, the "new" questionnaire. The methodological problem is not only the evaluation of the new items but also the evaluation of their effect, if any, on other continuing items in the survey. (See Gibson et al. (1978) for an example of how new questions added without formal advance testing affected estimates of traditional items in repetitive surveys.) By preserving the old questionnaire for most or part of the sample, comparisons with earlier data can still be made, although potentially larger sampling and nonsampling error may make trends more difficult to establish.[5]

---

[4]While infrequently used in Government research, such "piggybacking" is one cost-effective way to evaluate the wording, or alternative wordings, of a few key questions, especially those which designers think will not be affected by context.

[5]In addition to their testing function, split sample designs are also used in final questionnaire versions to minimize, or at least identify, biases associated with question or response category order. Occupational prestige studies conducted by the National Opinion Research Center provide an example of such use (e.g., Reiss, 1961: app. A).

Because of the design features of a split sample test, practical considerations of time, money, and other resources have frequently limited their use to surveys which will be unusually costly. Specifically, surveys with large sample sizes, panel designs, two-stage procedures requiring an initial screener survey, and surveys designed to be repetitive have typically been subject to more extensive testing than smaller, cross-sectional data collection efforts.

B. Method

In general, the requirements and procedures involved in different types of formal testing are similar. The description of the pilot study method contained in the first part of this chapter is applicable for split sample tests as well. Key differences introduced by the split sample design are discussed here.

1. Personnel and Skill Requirements

In some cases, more sophisticated (or simply different) statistical and analytical expertise may be required in split sample tests than in pilots. Data processing staff requirements (particularly editing and clerical coding procedures) may be less extensive, depending on the test design, although additional care must be taken in data processing not to accidentally distort test results. Similarly, a lesser or greater number of interviewers and field supervisors may be required, depending on sample design considerations.

2. Selection of Respondents

As in pilot tests, the sample of respondents must be selected by chance with known probabilities of selection, although probability and purposive strategies can be used in combination. Choice of an appropriate sample size for a split sample test depends on the aims of the test, the complexity of the sample design, the statistical techniques proposed for evaluation, and the degree of accuracy and confidence the designers require for the results. Although many surveys by Federal agencies are national in scope, split sample tests on a national scale--especially those involving face-to-face interviews--are often impractical because of constraints of funding, time, and the availability of field personnel.

A key aim of a split sample test may be the evaluation of questions aimed at a relatively small, but not geographically concentrated, subgroup--for example, households receiving Social Security retirement benefits--while also testing the questionnaire on all households. In effect, two separate tests could be fielded simultaneously using separate samples: one from Social Security Administration records and one from a general household list.

3. Preparation

In designing a split sample test, the following factors should be considered:

a. Specifying the test objectives

What is being tested? Why? What results are expected? How does the desired information fit into the overall purpose(s) of the survey? What is the test sample population? What key statistical tests are proposed to assess the results? What is the required reliability or accuracy for the statistics? Are statistical tests required for the total sample only? If not, which subgroups (e.g., Westerners, the self-employed, men) will be examined for response differences? What precision is required for subgroup tests? What criteria are to be used in deciding which questionnaire version is better?

In practice, specifying the answers to questions such as these is an iterative process that involves considerable coordination of effort among survey designers, statisticians familiar with sampling and survey research, data processing staff, and analysts who will be the primary users of the final survey data. For example, under specified assumptions about design effect and response levels, statisticians can provide estimates of sampling error for both the full test sample and for subgroups. However, statisticians need to know what size sampling error the analysts or sponsors are willing to accept for key estimates or statistics produced for the test. Statisticians usually work with and help the designers in clarifying data needs and in formulating such specifications.[6] During this process, budget and other resource limitations, including time for evaluation, may force compromises in goals.

b. Control procedures

In fielding the split sample test, procedures should ensure that the alternative treatments are administered randomly among respondents. The purpose is to decrease bias from factors other than those being tested, and it is frequently accomplished by means of instructions on cover sheets, odd-even numbered check digits or other identifiers that do not vary systematically among respondents. (When respondent Social Security numbers are available, the ninth digit can be used for up to 10 random assignments.) For personal interview surveys, assignments among interviewers also need to be controlled to avoid mistaking interviewer effects for differences in the alternative questions.

Control may be achieved more easily in telephone interviews because of opportunities for monitoring large numbers of interviewers on critical test items for relatively low cost. In either case, interviewer training should place considerable emphasis on asking test questions exactly as worded since interviewers, motivated as they are to encourage responses, may improvise in the field.

---

[6]Drawing on the Census Bureau's experience with survey sponsors, Cahoon et al. (1980) discuss the kind of information statisticians need, and need early, for the most efficient sample survey design. While focusing on the final survey, their paper may also be a useful reference to those planning split sample or other formal tests.

c. Processing plans

Processing data from split sample tests may sometimes be limited to hand or clerical tallies of only the items being tested; however, in most tests some, if not all, items are keyed and put through a computerized system. Regardless of the extensiveness of the system, the purpose is to avoid confounding the question test results with errors or alterations that may occur in subsequent stages. For that reason, split sample test data are frequently evaluated before any editing to ensure, as much as possible, that it is the questionnaire itself that is being evaluated. Later editing and/or imputation of the data may allow a second-stage evaluation.

d. Analysis plans

Statistical tests planned for the evaluation of test results should be specified in advance. Time constraints following the test should also be considered in the planning stage. Since test results often need to be turned around quickly to make decisions for the final survey or the next questionnaire testing stage, such constraints may preclude extensive analysis or suggest modification of optimal statistical approaches.

To evaluate the test results, numerous statistical approaches are possible. First, designers may look at response distributions on single items. For split sample questions, the issue may be whether an item results in significantly more nonresponse in one treatment than another. Or, more respondents may report a given behavior in one question version than they do in another.

Another area of investigation is the relationship among test items in split samples and other items like sex, age, education, or other characteristics. For such investigations, statistical tests can determine whether associations between the test items and another characteristic(s) in two (or more) treatments are significantly different. Various methodologies are available; the key emphasis here is that the analytical methods should be anticipated in planning the test.

For some types of items, data may be available from other sources allowing treatment responses to be compared to other data. For example, the number of persons receiving Social Security or voting in a Presidential election is known, and survey responses to questions about such items can be benchmarked against independent estimates from outside sources to establish the general validity of the responses. In other cases, consistency or validity can be directly established for individual cases by comparison with information in administrative records (see Chapter 10 for a description of record checks). Reinterviews with the same respondents conducted shortly after the test have also been used as reliability checks.

However, for many question types--especially questions seeking attitudes, values, or opinions--the "true" value is unknown, although models of "construct validity" have been established to examine the degree to which

an attitude question measures what it is supposed to measure.[7] Other methodological research has utilized replication studies to test the reliability of results of attitude surveys (e.g., Schuman and Presser, 1981). Unfortunately, these options are rarely available in split panel tests used for questionnaire design, and researchers generally must be guided by theory, experience, or intuition. That is, if two treatments aimed at measuring the same thing produce significantly different results, it frequently remains a matter of informed judgment as to which result is the more reliable or valid one. "More is better" (e.g., more reported income) is a frequent rule of thumb in comparing treatment responses, but one which may not be correct.

e. Observation feedback system

In addition to the formal statistical analysis, subjective evaluation techniques (e.g., observation of interviews, interviewer debriefing) can be employed in a split sample test. Information gained from these methods may help interpret statistically significant differences or unexpected results found after the test is complete.

4. Operation

The data collection phase of the test can proceed in the same way as a pilot study, but with the additional constraint that monitoring should assure the proper correspondence between sample cases and the appropriate treatment group. Compared with the pilot, the split sample data processing phase may be less extensive and data analysis more concentrated on the test items and their potential impact on other variables.

5. Time Considerations

The amount of time required to plan, execute, and analyze the results of a split sample test is usually similar to that of a pilot study. However, it may be distributed somewhat differently, possibly requiring, for example, more advance planning but little or no editing or clerical coding. See section B-5 on time requirements of a pilot for a detailed discussion of this topic.

6. Cost Considerations

The cost factors outlined in section B-6 on pilot studies are also appropriate for a split sample test.

7. Mode of Data Collection

Split sample tests are suitable for any data collection mode. Recent advances in telephone interviewing--especially computer-assisted telephone

---

[7]Construct validity, as described by Andrews and Withey (1976:182), refers to "the relationship of an observed measure to a theoretical construct (or concept)." Since modeling such validity estimates involves unmeasured variables, investigations rest on theoretical assumptions about the relationships among unobserved and observed variables.

interviewing (CATI), sometimes combined with random digit dialing (RDD)--may encourage the use of split sample testing for several reasons. Telephone interviewing is generally cheaper and faster than other interviewing modes. In addition, the ability of a CATI system to randomly select alternative question wordings or question sequences for each interview eliminates many of the operational difficulties in conducting split sample tests using hard copy questionnaires.

C. Examples

1. Example 1: The 1979 Income Survey Development Program's Test of Attitude Measures

a. Introduction

The Income Survey Development Program (ISDP) was established in 1976 to develop and test procedures to improve survey data on income, on participation in government aid programs, and economic well-being.[8] Because of known measurement problems and because results were to be used in a series of national panel studies, the testing phase was considerably more extensive than is usual for household surveys. The program was jointly sponsored by the Department of Health and Human Services and the Bureau of the Census.

The 1979 ISDP Research Panel included a number of split sample and other tests. A single example--a test of two alternative subjective measures of well-being--is described here. This example was chosen because its straightforward field procedures are easily transferable to many survey situations and because the evaluation incorporated several common techniques.

b. The problem

Attitudinal measures originally developed and tested by Andrews and Withey (1976)[9] had been used in earlier ISDP field tests (Vaughan and Lancaster, 1979, 1980). The items asked respondents to rate their life as a whole, their personal economic situation and, for those with children, their income in terms of providing for their children. The items were designed to provide an additional means of evaluating the impact of government aid programs and to assess overall economic well-being.

Previously, respondents answered by choosing one of seven labelled categories as shown in the left panel of Figure 4. Results using these seven

---

[8]For further description of ISDP goals and activities, see Ycas and Lininger (1981). For detailed documentation of the 1979 ISDP Research Panel, see "Income Survey ..." (1982). Material summarized in this example is drawn primarily from Olson and Vaughan (1982).

[9]To examine well-being, Andrews and Withey tested five measurement methods and evaluated them using four criteria: construct validity, distribution form, category labeling, and ease of use. They found the delighted-terrible scale to be methodologically superior, especially in terms of validity, but weaker than some other measures in distribution form.

"delighted-to-terrible" categories showed that reported attitudes have a strong positive skew, with most responses clustering on the "delighted" end of the scale. Empirically, such skewed distributions and the lack of variation hampered many applications of the scale, especially in multivariate analyses.

c. Design of the test

Because of these limitations, additional response categories were developed. The result was a 10-category version of the "delighted-terrible" scale which is shown in the right panel of Figure 4. This expanded set of response categories was primarily meant to allow respondents more choice among the positive categories. Designers were uncertain, however, whether respondents could make meaningful distinctions among so many items.

Figure 4. The "Delighted-Terrible" Response Categories

Delighted	Delighted
	Very pleased
Pleased	Pleased
Mostly satisfied	Mostly satisfied
	Somewhat satisfied
Mixed (about equally satisfied and dissatisfied)	Mixed (about equally satisfied and dissatisfied)
	Somewhat dissatisfied
Mostly dissatisfied	Mostly dissatisfied
Unhappy	Unhappy
Terrible	Terrible

Therefore, it was decided to test the items using a split sample aimed at assessing whether a greater proportion of valid variance (in the sense of meaningful distinctions) was captured in the 10-item scale than in the 7-item one.

d. Field implementation[10]

The 1979 panel involved a national probability sample of 7,500 households in which all adults were to be personally interviewed. The sample size was dictated by the nature of the larger ISDP mandate and was far larger than

[10]Census interviewers conducted all interviews. Although there were six interviews with the same respondents in the 1979 ISDP Panel and the experiment was repeated at the conclusion of each of the first three interviews, the test described here used data from the initial interview only. It might also be noted that the 1979 Panel included samples from two administrative lists; those respondents are excluded from this analysis.

necessary for evaluating the single test presented here. Readers should bear that caveat in mind.

Sample households were divided into random halves prior to interviewing, and a numerical designation indicated the half to which each household was assigned. Since the questions are attitudinal ones, interviewers were instructed to ask them only of adults interviewed personally (see check items S1 and S2 in Figure 5).[11] While all respondents were asked the same questions, half of the households received the seven- and the other half received the ten-category response choices. (See check item S3 in Figure 5.)

Flashcards listing the "delighted-terrible" response categories were used for the two sets of questions; interviewers were instructed to read the questions exactly as worded, and not to read the answer categories unless respondents were blind or unable to read. If a respondent was unsure of which of two or three boxes to choose, interviewers were to probe by saying that "the one that comes closest to the way you feel" be chosen. Finally, interviewer manuals emphasized the importance of neutrality and accuracy in administering these attitudinal items.

e. Field evaluation

Staff researchers and questionnaire designers observed as many interviews as possible. Respondents (and interviewers) appeared to enjoy the opportunity to express their attitudes, and respondents did not appear confused by the longer list. Written observation reports and informal discussions were used to elicit observers' views about the questionnaire and interview interaction. Field observers noted that the presence of another person--especially a spouse--during the questioning may have influenced the answers given by some respondents.[12]

---

[11]Although ISDP households were all chosen according to probability designs and the test was administered on a probabilistic basis, respondents to the attitudinal questions depended on who acted as self-respondents. In the 1979 Panel, rules governing situations in which a proxy could be accepted were also being tested. For one-third of the sample households with very demanding self-response rules, a self-response rate of about 90 percent was obtained. For the remaining two-thirds with less demanding self-response rules, approximately 70 percent were self-respondents.

[12]While potentially related to the test items, statistical evaluation assumed that any effects in the first interviewing wave were randomly distributed among treatment groups. The observation, however, led to the inclusion of an item in a subsequent wave to allow the interviewer to record such situations so that analysts could assess the effects, if any, that the presence of another person had on respondents' expressed attitudes.

Figure 5. Attitude Test Items, 1979 Income Survey Development Program, Wave 1*

CHECK ITEM S1 ▶ Is . . . a self-respondent?	(4415) 1 ☐ Yes — *Go to Check Item S2* 2 ☐ No — *Go to Check Item S6*
CHECK ITEM S2 ▶ Is . . . under age 21 and living with parents?	(4416) 1 ☐ Yes — *Go to Check Item S6* 2 ☐ No — *Read statement and go to Check Item S3*
READ: **We have now completed the questions that deal with the facts of your financial situation. As I mentioned at the beginning of the interview, the last few questions are about how you actually FEEL about your financial situation and how things seem to be going for you these days in general.**	
CHECK ITEM S3 ▶ Is the last digit of the serial number even?	(4417) 1 ☐ Yes — *Ask 1* 2 ☐ No — *SKIP to 4*
Hand respondent flashcard G. **1. The first questions can be answered by telling me what word on this card comes closest to how you feel: "delighted," "mixed," and so forth on down to "terrible."** **Please tell me the feeling you have now, taking into account what has happened in the last few months and what you expect to happen in the near future.** **The first question is quite general. How would you say you feel about your life as a whole?**	(4418) 21 ☐ Delighted 22 ☐ Very pleased 23 ☐ Pleased 24 ☐ Mostly satisfied 25 ☐ Somewhat satisfied 26 ☐ Mixed (about equally satisfied and dissatisfied) 27 ☐ Somewhat dissatisfied 28 ☐ Mostly dissatisfied 29 ☐ Unhappy 30 ☐ Terrible
2. The next few questions are about your income. Overall how do you feel about your (family) income; that is, all the money that comes in to you (and other members of your family living with you)?	(4419) 21 ☐ Delighted 22 ☐ Very pleased 23 ☐ Pleased 24 ☐ Mostly satisfied 25 ☐ Somewhat satisfied 26 ☐ Mixed (about equally satisfied and dissatisfied) 27 ☐ Somewhat dissatisfied 28 ☐ Mostly dissatisfied 29 ☐ Unhappy 30 ☐ Terrible
CHECK ITEM S4 ▶ Is . . . responsible for children living in the household (parent, guardian, etc.)?	(4420) 1 ☐ Yes — *Ask 3* 2 ☐ No — *SKIP to Check Item S6*
3. How do you feel about your family's income in terms of being able to provide the things you think the children should have?	(4421) 21 ☐ Delighted 22 ☐ Very pleased 23 ☐ Pleased 24 ☐ Mostly satisfied 25 ☐ Somewhat satisfied 26 ☐ Mixed (about equally satisfied and dissatisfied) 27 ☐ Somewhat dissatisfied 28 ☐ Mostly dissatisfied 29 ☐ Unhappy 30 ☐ Terrible } *SKIP to Check Item S6*
Hand respondent flashcard H. **4. The first questions can be answered by telling me what word on this card comes closest to how you feel: "delighted," "mixed," and so forth on down to "terrible."** **Please tell me the feeling you have now, taking into account what has happened in the last few months and what you expect to happen in the near future.** **The first question is quite general. How would you say you feel about your life as a whole?**	(4422) 1 ☐ Delighted 2 ☐ Pleased 3 ☐ Mostly satisfied 4 ☐ Mixed (about equally satisfied and dissatisfied) 5 ☐ Mostly dissatisfied 6 ☐ Unhappy 7 ☐ Terrible

*Questions 5 and 6 (not shown) repeat the question wording of items 2 and 3, but use the seven response categories of item 4.

f. Evaluation[13]

First, item nonresponse associated with the two scales was examined. It was thought that nonresponse on the experimental 10-point scale might be higher if respondents found it too difficult to discriminate among so many categories. However, results showed that item nonresponse rates were relatively low, ranging from .5 to 5 percent, and respondents using the 10-point scale were as likely to respond as those using the 7-point scale.

Frequency distributions on the two scales for the three questions are presented in the upper panel of Table 1; summary statistics, using numbers arbitrarily assigned from 1 to 7 and 1 to 10, are provided in the lower panel. Overall, the data suggest that the 10-point scale resulted in somewhat more dispersion and lesser positive skew than the 7-point scale. For example, a lower percentage of respondents chose one of the two most positive categories in the 10-point scale, and positive skew was reduced for all three test items (reductions of about 40 percent occurred for the income assessment items).

Variation in respondents' subjective assessments of well-being was then related to their objective characteristics as reported in the survey. Bivariate associations between attitudes--especially individuals' assessments of income--and income showed the expected relatively high correlations. However, the results also showed the 7-point scale to be as strongly associated with income as the 10-point scale, suggesting that the larger variance yielded by the 10-point scale might not be meaningful.

To further explore that question, a simple multivariate model, regressing income on the "income adequacy for children" attitude item and controlling for family size, was used. Under selected specifications of measured income, consistently more variance was explained in the regressions using the 10-point dependent variable than in those using the 7-point measure, although in two regressions, estimated with an income variable believed to be "weak," differences of only 8 percent were found.

For the most part, however, the regression models showed encouraging relative differences in explained variance using the 10- versus the 7-point scales. To date, however, statistical evaluation has not provided an unequivocal answer to the issue of construct validity. Work in this area is continuing and more conclusive results in the future may lead to a clearer recommendation about the use of these items in future questionnaires.

---

[13]Denton Vaughan, of the ISDP staff (currently with the Social Security Administration) designed the experiment and conducted the evaluation which drew heavily on the work of Andrews and Withey (1976) and Atkinson (1977). Readers interested in this experiment on the measurement of well-being may wish to look at the data themselves. Public use data tapes from the 1979 ISDP Panel are available from the National Technical Information Service.

Table 1. Distribution of Responses to Three Test Questions Using 7- and 10-Point Scales

Category	Item and number of scale points					
	Life in general		Family income overall		Family income for children	
	10-point scale	7-point scale	10-point scale	7-point scale	10-point scale	7-point scale
	Part 1. Percent Distributions					
Total	100	100	100	100	100	100
Delighted	9	11	2	2	2	4
Very pleased	15	-	5	-	7	-
Pleased	21	29	14	16	14	16
Mostly satisfied	23	34	21	33	19	31
Somewhat satisfied	8	-	14	-	12	-
Mixed	11	17	13	23	13	23
Somewhat dissatisfied	4	-	12	-	14	-
Mostly dissatisfied	3	5	7	13	8	12
Unhappy	3	2	5	7	5	8
Terrible	2	3	6	6	6	6
	Part 2. Summary Statistics					
Mean	4.0	2.9	5.4	3.7	5.4	3.7
Standard deviation	2.1	1.3	2.3	1.4	2.3	1.5
Percent in highest category	9.1	10.6	2.3	1.9	2.0	3.6
Percent in two highest categories	24.0	39.5	6.8	18.0	8.9	19.9
Percent below "mixed"	12.2	9.8	30.5	26.0	33.8	25.9
Skew	.8	.9	.4	.6	.3	.5
Kurtosis	.3	1.1	-.6	-.2	-.8	-.3
Coefficient of variation	52.1	44.6	41.7	37.8	42.6	39.6
Number of cases	5,753	5,458	5,741	5,467	2,460	2,276

Note: For the 10-point scale, assigned numerical values ranged from 1 (delighted) to 10 (terrible). For the 7-point scale, values ranged from 1 (delighted) to 7 (terrible). Distributions are based on weighted counts.

2. Example 2: National Center for Health Statistics

a. Introduction

The National Center for Health Statistics (NCHS) has undertaken numerous studies to improve the reporting of health events in household surveys.[14] The example of a split sample test presented here was selected for three key reasons. First, it illustrates the use of a small and unusually homogeneous sample, showing the strengths and weaknesses of such an approach. Second, it tested three questionnaires reflecting different strategies of questionnaire design. Third, it illustrates the successful application of hypotheses developed in another field--cognitive psychology--to survey research.

The concrete problem facing the designers of this test was the underreporting of their key dependent variable, "health events" in a given time period (e.g., the number of dental visits in the last 14 days). Especially likely to be underreported were health conditions of low impact to the respondent and those occurring considerably prior to the interview.

The test was developed using a cognitive model of how people learn, store, and retrieve information. Methodologically, the aim was to determine whether reporting can be significantly increased by focusing on and aiding the recall tasks facing respondents. The model, described in Cannell et al. (1977:52-54), suggests that an event is reported only if the researcher can design a survey question/stimulus that can spark the memory during the interview. For example, a single event--number of dental visits--may be recalled by the respondent in terms of money, pain, or lost work time, and a direct question on dental visits may not get an accurate answer.

b. The questionnaire

To test some hypotheses generated by the model, three questionnaires were developed for a split sample test: an extensive questionnaire, a diary with a follow-up procedure, and a control questionnaire. All relied on personal interviews, although the diary follow-up was partially self-administered.

In the extensive questionnaire there were many questions aimed at providing respondents "with multiple and overlapping frames of reference and cues." The strategy rested on the assumption that respondents could more easily recall health conditions through "some specific behavior implications" (e.g., activity restrictions, medicines, diet, visits to doctors) than through a conceptual or general framework (Laurent et al., 1972:3). For example, previous field work showed that questions about operations usually resulted in reports of major surgery, but questions about stitches elicited reports of minor surgery as well. Therefore, standard questions included additional probes, and general medical terms as well as more popular language were used. Finally, the pace of the interview was designed to be more relaxed by allowing more time for recollection and reporting and by the use of transitions between sections.

---

[14] The example summarized here was adopted from material in Laurent et al. (1972) and Cannell et al. (1977).

A "body review" of aches and pains and a series of questions on symptoms (e.g., "Have you had any pain or soreness in your joints?") opened the extensive interview. When respondents reported a symptom, interviewers asked "Do you have any idea what causes it?" in an attempt to help the respondents better define and isolate the underlying health condition. Next, questions about the respondents' medical history specified various time dimensions (e.g., childhood, last week) as another approach to uncovering events stored in the memory. Behavioral implications were referenced in the next questions. Two checklists of chronic conditions provided a direct items-recognition approach to conclude the interview.

Reviews of previous research on health diaries and informal tests of various procedures led to the second experimental approach. It utilized a diary procedure, in which the respondent kept a health record every day for a week in an 8-page booklet containing seven simple questions on health events. A short personal interview took place at the end of the diary week. The design operationalized two major ideas:

> The first was to facilitate the respondent's task of remembering, by minimizing the period of time between the event and its solicited recall....The second idea was to consider this daily recording activity as a sensitization device for health thinking and reporting, which could result in increasing the reporting level in a follow-up interview (Laurent et al., 1972:5).

In the follow-up interview, each diary question was carefully reviewed, answers were clarified when necessary, and a short structured set of questions--the chronic conditions lists and items on present effects of past accidents, injuries, dental visits, and hospitalizations--were asked.

The control questionnaire used a single direct question for obtaining information on each major health item. This short questionnaire was comparable, though not identical, to that used in the current National Health Interview Survey (NHIS). To sensitize the respondent, the interview opened with a checklist of 19 symptoms. Questions were then asked on recent health events, including restrictions of activity, and on present effects of past injuries or illnesses. Then, the chronic conditions checklists, identical to those used in the two experimental questionnaires, were administered. The interview concluded with questions about recent visits to the doctor and hospitalizations or dental visits in the past year.

In addition to the chronic conditions checklist, items on hospitalizations, dental visits, demographic characteristics, and a general health rating were identically worded in all three approaches. Other questions were similarly worded across the three instruments. Then, at the conclusion of the health questions in every interview, interviewers asked a standard series of questions about each reported condition. The resulting "condition table" was designed to separately record the first report of any health problem mentioned by respondents. The purpose of these standardized questions was to allow

> comparative evaluation of the three experimental collection methods through an analysis of the impact nature of the

information reported. This was designed to test the idea that attempts to facilitate recall could accomplish their mission by eliciting lower impact information that is commonly underreported (Laurent et al., 1972:6).

c. Sample design

Previous research on health reporting had shown that "characteristics of the respondent are not nearly as consistent, nor as strong in their influence on underreporting, as are characteristics of the event" (Cannell et al., 1972:16). For this reason, and because of the experimental nature of this test and the desire to minimize costs, a geographically concentrated and relatively homogeneous sample was selected. Specifically, all cases were in Detroit, and a modified area probability sample with clustering was used to locate "low-middle and middle socioeconomic groups, English-speaking, native-born, white females between 18 and 65 years of age." The three questionnaires were randomly assigned to households within each sample block (Laurent et al., 1972:9-10). The design yielded 462 occupied dwelling units, containing 356 dwellings with eligible respondents. Only 1 respondent per unit was interviewed, and 305 interviews were completed.

d. Training and field operation

Under contract to NCHS, the Survey Research Center (SRC) at the University of Michigan's Institute for Social Research conducted the test using six interviewers from the SRC staff. Interviewer training was extensive, lasting 2 weeks, and included role-playing, practice interviews in the classroom and the field, and feedback procedures. The actual interviews occurred between April and June 1968.[15]

e. Evaluation

Because underreporting of health events was a known problem, comparisons among the questionnaires focused on the amount of reported information. The assumption was that the more health information reported the better; no outside records were used for validation. There were two main types of dependent variables: (1) the number of health conditions reported and (2) the impact level (i.e., the amount of medical care, psychological concern, and other indexes of salience to the respondent) of reported information.

The overall response rate of 88 percent was quite similar among the three questionnaires. Demographic characteristics of respondents were also similar with the exception of education, which was highest in the diary group. However, correlations within the treatments between education and the key dependent variables were not statistically significant.[16]

---

[15]Editing and coding were unusually complicated procedures in this test. See Laurent et al., 1972:10 for a description of these operations.

[16]While not technically correct, assumptions of simple random sampling were used for the analysis. However, the authors note that using the same area and the same design, another study had shown an average design effect of 1.03 times simple random sampling.

The evaluation first examined the mean number of health conditions reported per person in each of the three questionnaires. As seen in Table 2, results supported the hypothesis that the multistimuli approach of the extensive interview increased reporting: the 7.9 reported conditions in the extensive interview were significantly greater than the 5.1 reported in the diary or the 4.4 in the control.[17] The hypothesis that diaries would also increase reporting received less clear support. The difference between reported health conditions in the diary and control questionnaires was statistically significant only at about the 10-percent level.

To learn more about the source of these differences, conditions were classified into five types, and Table 2 also shows the number of reports of each type, by questionnaire version. Again, the extensive questionnaire achieved higher reporting than the control among all types, although, as the authors note,

> whenever the control questionnaire uses an extensive recognition type of approach, such as the recognition lists of chronic conditions, a reduction of the gap between the two techniques can be observed. An increase in the amount of information reported still exists in the extensive technique but is no longer statistically significant (Laurent et al., 1972:16-17).

Compared with the diary follow-up interview approach, the extensive questionnaire also achieved higher reporting except for acute conditions. This particular strength of the diary procedure was expected, but since the reporting of chronic conditions did not significantly differ from reports in the control interview, doubts were raised about the general sensitization function of the diary.

When reported conditions were dichotomized into those first noticed less than 3 months ago and those first noticed 3 months ago or longer, reporting of both recent and older conditions was significantly higher with the extensive questionnaire than with the control questionnaire. But compared with the diary follow-up questionnaire, the extensive questionnaire got significantly higher reporting only for longer term conditions (Laurent et al., 1972:19). As the authors observe, these results are not surprising since older reported conditions are more likely to be chronic and recent reported conditions more likely to be acute.

The second key dependent variable was the level of impact on the respondent of the reported health conditions. It was hypothesized that the extensive and diary follow-up questionnaires would improve reporting of low impact conditions but have little, if any, effect on high impact reporting. Thus, the predicted result was a lower mean level of impact reported using these questionnaires compared with the control. For testing the hypothesis, an impact level was constructed for every eligible condition using, for example, evidence of frequency (or levels) of discussion with doctors, medications

---

[17] Student's tests were used throughout the analysis to evaluate the significance of differences between means.

taken, days in bed, and pain (Laurent et al., 1972:24-30). Results supported the hypothesis and further showed that the extensive questionnaire produced more complete reporting of serious (i.e., high impact) conditions. Differences among the questionnaires in mean level of impact according to whether conditions were chronic or acute were also uncovered.

Table 2. Mean Number of Conditions Reported Per Person, by Condition and Collection Procedure

Reporting variable	Collection procedure					
	Extensive	Control	Diary	Extensive-control	Diary-control	Extensive-diary
	Mean number of conditions per person			Difference between means		
Total	7.88	4.42	5.08	1/3.46	0.66	1/2.80
Chronic conditions on recognition lists	3.54	3.25	3.29	.29	.04	.25
Other chronic conditions	2.75	.74	.58	1/2.01	-.16	1/2.17
Illnesses in last 14 days	.58	.28	2/.69	1/.30	1/.41	-.11
Injuries in last 14 days	.24	.05	2/.30	1/.19	1/.25	-.06
Other unclassified conditions	.76	.10	.22	1/.66	.12	1/.54

1/ $p \leq .01$

2/ These figures in diary technique refer only to the last 7-day period, a restriction which enhances the observed differences between diary and the other techniques.

Source: Laurent et al., 1972:16.

f. Summary

By emphasizing various ways of encouraging respondents to recall health events, this small test produced extremely encouraging results. The extensive questionnaire with multiple probes and cues significantly increased reporting in all groups of health conditions. Compared with the control, the extensive and diary questionnaires also produced higher reporting of health conditions of low impact to the respondent. The diary follow-up procedure resulted in more reported acute conditions, although hypotheses about the sensitization function of the diary were not generally supported.

Because of the special demographic characteristics of the sample, generalizing the results to other groups cannot be done with any certainty. The test was instead part of a larger and long-term research effort aimed at achieving greater understanding of survey techniques for better reporting of health events.

Methodologically, the improved reporting was "interpreted as the result of a greater correspondence between the questioning procedures and the manner in which respondents organize health information in memory" (Cannell et al., 1977:59), although the authors caution that motivational factors were not controlled in the study. Rather, "the major outcome was a pragmatic one; techniques designed in a cognitive framework to facilitate recall have proved effective in increasing reported information" (Cannell et al., 1977:60).

Part IV

Techniques for Evaluating the Questionnaire Draft

Part III of this report discussed various types of testing that are conducted to examine a questionnaire under field conditions. In Part IV, a number of techniques are presented for evaluating the performance of questionnaires. Some are routinely used in conjunction with testing; others are used less frequently or possibly altogether independently of other testing. These techniques use a variety of sources to evaluate the questionnaire, such as interviewers, observers or independent records.

The first technique, discussed in Chapter 7, is a way to determine the respondent's frame of reference at the time of the interview. By probing to ascertain the meaning that certain words, phrases, or situations may have for different respondents, the extent to which they are understood in the manner intended by the questionnaire designers can be evaluated. This technique can be incorporated into any stage of questionnaire development from initial informal testing through the actual survey administration. The "evaluator" in this technique is the same person who provides the original response, but the evaluation response is made to a different stimulus.

In Chapter 8, the contributions made by professional staff in the role of observer/monitor of the interviewing process are described. Generally, interviews conducted in person are observed and those conducted on the telephone are monitored as part of both informal and formal test evaluations.

This technique can provide valuable insights into problems with a draft questionnaire during the developmental stages or with a questionnaire used in repetitive surveys which needs to be changed.

The assessments of interviewers are also useful tools in evaluating questionnaires. Chapter 9 presents descriptions of two methods which can be used to elicit interviewer judgments concerning how well a questionnaire achieves its objectives. The first method, interviewer debriefing, consists of verbal discussions on aspects of the questionnaire which relate to the quality of the data collected. Such debriefings, like observations by professional staff members, are a routine component of the evaluation of informal and formal tests; they can also contribute to questionnaire development for successive interviews in repetitive surveys.

The second type of interviewer assessment is collected in a more systematic fashion. In structured post-interview evaluations, questionnaires requesting information about the interview situation or the respondent's attitude toward or participation in the survey are administered to the interviewers. This technique is used less frequently than some of the others described here. It is usually undertaken in conjunction with formal tests or repetitive surveys as opposed to informal tests, and can be helpful in interpreting survey findings.

Validation of data collected in a survey with comparable information on the same persons from independent records constitutes another type of evaluation technique described in this section. Record checks, described in Chapter 10, are generally conducted as separate field tests rather than as part of the tests described in Part III. They are a particularly useful tool for dealing with subject matter where there is relatively little survey experience to draw upon; however, their use is limited by the availability of records on that topic.

The final chapter (Chapter 11) describes a technique, which like the one presented in Chapter 7, obtains evaluation information from respondents. This technique is called a response analysis survey. Evaluation information is obtained at a different time than the survey data and using a different method--a personal visit follow-up interview is generally conducted after respondents have completed a mail questionnaire. It is particularly useful in the evaluation of questionnaires for which accurate response depends heavily on administrative or other kinds of records.

Although not the subject of detailed discussion in this report, several other evaluation techniques deserve brief mention here. Just as observation and interviewer debriefing are routinely incorporated into informal tests, statistical analytical techniques are routinely employed in the evaluation of formal tests. Illustrations of the use of statistical tests in making questionnaire design decisions are contained in the examples in Chapter 6 on formal tests.

Reinterview, a technique which is commonly used as a quality control device, is also used to measure the reliability of survey responses. It requires that a sample of survey respondents be recontacted and asked either the same questions as those contained in the original interview or different questions

designed to yield the same information. In some instances, this technique can also be used to elicit information about questionnaire design--for example, the 1970 census included a question about language spoken in the home. The degree of inconsistency between the responses to the original interview and the reinterview was quite high, suggesting that there was a problem with this question. In the 1980 census, the question was deleted.

Structured post-interview evaluation by interviewers, as described in Chapter 9, is a technique which can be extended to respondents as well. Conducting post-interview evaluations with respondents as described here differs from reinterview. The post-interview questions are not designed to elicit information about the survey content, but instead ask about the respondents' attitudes or behaviors relevant to the interview situation itself. For example, how much did respondents know about the purpose for which the data were collected? Were the respondents able (i.e., did they have sufficient knowledge or information) to answer the questions they were asked? Answers to such questions provide indirect measures of the validity of the data collected. If respondents do not have an understanding of the uses to which the data will be put, they may expend less energy in trying to provide accurate information. If the respondents do not know the answers to the questions which they are asked, the answers they provide (and they may provide answers rather than appear ignorant) will be of questionable value. An illustration of the use of respondent post-interview evaluations in conjunction with interviewer post-interview evaluations is described in the examples in Chapter 9.

Chapter 7
Investigating Respondents' Interpretations of Survey Questions

I. INTRODUCTION

One way to evaluate questionnaires is to investigate respondents' understanding of the intent of specific questions and the meaning of their replies to those questions. This technique is called frame-of-reference probing, and is done by asking the repondent some additional questions.[1] It is designed to address concerns about whether the questions, definitions, and instructions proposed for a questionnaire convey the frame of reference desired. Probing to determine respondent frame of reference can be especially useful when words (like "crime") that are key elements in a survey are thought to carry emotional impact.

The probing questions can take different forms: either ad hoc questioning by the interviewer or administration of a set of questions written in advance (called "structured"). Ad hoc questioning usually takes place after the survey questionnaire has been administered. When structured follow-up questions are asked, the probing might be done immediately after the question containing the word or concept of interest is asked; alternatively, it might be done after the survey questionnaire has been completed.

Frame-of-reference probing can be incorporated at various stages of the questionnaire development process. It might be planned as a part of a pilot study or field test or it might be done during the actual survey.

II. METHOD

A. Personnel and Skill Requirements

This technique is implemented by interviewers, and to some extent, the skill requirements involved depend on whether the probing takes the form of structured follow-up questions or unstructured questioning. In the former case,

[1] This technique has been used extensively in England by William Belson (1981).

Principal Contributor: Anitra Rustemeyer Streett

regular interviewing skills are required; in the latter, more extensive interviewing skills such as detailed probing, ability to think quickly, and others described in Chapter 2 on unstructured interviewing are also necessary.

When this technique is used during informal testing, it may be preferable for researchers/questionnaire designers to conduct the interviews to give more insight into respondents' interpretations of the word or phrase of interest.

B. Selection of Respondents

The way respondents are selected for frame-of-reference probing depends on which stage in the questionnaire design process the method is used. During the questionnaire development process, respondents are selected using the same purposive selection strategies as those used in informal tests or unstructured interviews. If respondents' interpretations of questions in formal tests or actual surveys are subjected to investigation using this technique, however, respondents have already been selected through scientific procedures. Depending on time and resource constraints, everyone in the sample can be included in the frame-of-reference probing, or respondents can be subsampled and the additional probing questions asked of only a percentage.

C. Preparation

In advance of data collection, the following basic decisions need to be made:

1. Decide when during the questionnaire design process to probe respondents' interpretations.

During the questionnaire development, probing to determine respondent frame of reference for key concepts can facilitate improvements in question wording and thereby avoid collecting data that cannot be properly analyzed statistically. This type of question investigation can warn the survey designer of ambiguities that will cause respondent confusion and irritation. If ambiguities concerning the meaning of questions are present, it is likely that the interviewers will be asked to explain what is meant or what type of answer is wanted. When interviewers are asked to explain questions, the chance of interviewer bias increases dramatically.

If probing to determine respondent frame of reference is included in the final questionnaire used for the survey, it can help to illuminate the answers provided in the survey. The answers to the probing questions may help the survey analyst to understand what appear to be inconsistent answers. And in a repetitive survey, problem questions can be deleted or changed for subsequent interviews.

2. Decide which words or phrases to probe.

Words or phrases that are central to collecting uniform information and thought to be susceptible to misinterpretation should be subjected to investigation. For example, in a study designed to evaluate the seriousness of various crimes, the respondent might be asked to rate the seriousness of an event described as "An offender injures a victim and the victim dies." To

know whether the respondent answered in general terms or attributed specific circumstances to the event before rating its seriousness, additional probing should be done to determine how each respondent interpreted the question.

3. Decide where in the interview to probe.

If the questions added for the frame-of-reference probing do not disrupt the interview (by changing the subject, for example) and are not expected to bias the remaining survey questions, then it is probably best to ask them immediately after the question where the word or phrase of interest appears. By placing the probing questions immediately after the survey questions of interest, there should be no doubt as to what word or phrase is being referenced. If the probing questions might disrupt or bias the interview (such as detailed questions about sources of income, traffic accidents, or the nature of mental illnesses in the family), those questions could be placed near the end of the interview and preceded with a transition statement such as "Earlier I asked you about ...; now I have just a few more questions about that."

4. Arrange probing so that only a few questions (two to four) are probed with a respondent.

The number of survey questions to be investigated by frame-of-reference probing is decided by the researcher. However, if more than about two to four words or phrases are to be investigated, it might be better to limit the number subjected to probing with any one respondent and interview a larger number of persons to collect enough data. Important considerations in setting the number of questions to be investigated are the total length of the interview and the respondent's tolerance for being questioned in detail on subjects for which (s)he may have little interest and/or knowledge. Unless the respondents selected for this type of interviewing are known to be especially knowledgeable or interested in the topics to be probed, it may be best to assume a low level of knowledge and interest and arrange the probing questions accordingly.

5. Determine how many and what kind of probes to use to investigate each word or phrase under study.

The optimal number of questions used to determine the meaning attached to a word or phrase is probably about three to five. If too few probes are used, there is the risk of superficial or inadequate treatment of the subject; if too many are used, there is the risk of being tedious, appearing to challenge or question a respondent's views, beliefs, or attitudes, or of appearing to be administering a test in which there are "right" and "wrong" answers.

Clearly, adding questions to an interview results in a more time-consuming interview. In addition, there may be some respondents who will dislike being asked to report information such as what things they consider to be ..., what they were thinking about when they answered a question, or other questions requesting them to think about how they think about things. If the probing questions are carefully worded, it should be possible to avoid putting respondents "on the spot." An illustration of a question that was carefully constructed to avoid putting a person "on the spot" is: "Speaking of crime,

everyone agrees some acts are crimes, but there are different ideas about others. Do you believe it is a crime for someone to ...?"

6. Arrange method of probing and presentation of additional questions.

The method of probing depends on the stage of the questionnaire design process at which the technique is used. When it is used for questionnaire development, it might be more useful to the researcher if interviewers are given guidance on what information is desired and then allowed to develop their own follow-up questions. To some extent, the choice between structured and unstructured methods during developmental work depends on the level of experience of the interviewers; less experienced interviewers and those not familiar with research methods may require more structured assignments.

If used during the survey itself and if all respondents are to be asked all frame-of-reference probing questions, the follow-up questions should be printed on the questionnaire so that they will be asked in the same way, and at the same time during the interview, of all respondents.

7. Establish system to record results of the probe.

Two common ways of recording results of unstructured interviewing are tape recording and having a second person accompany the interviewer to take notes. If structured questions are used (with printed questions to be read), then precoded answers may be developed to aid the interviewer in recording the answers quickly.

8. Develop technique for reconciling survey question response with probing response if the two answers are expected to be the same.

Sometimes the frame-of-reference probing questions ask for the same type of information as the survey question, but in a different manner. When the same type of information is asked, the respondent may seem to give quite different or contradictory responses to the frame-of-reference probing than (s)he did to the survey question. Reconciliation of responses is important for these cases. If this happens, the interviewer might say, "In light of what you've just been saying, I'd like to go back and ask again one of my earlier questions; ...(repeat question)."

D. Operation

Since frame-of-reference probing is generally done in conjunction with one of the stages of testing or with the survey itself, the selection of a site and other operational details are taken care of in planning for the main event. Some additional details may be necessary to accommodate the use of this technique, however. For example, if experienced interviewers rather than researchers are involved, they may require extra training on how to ask the additional questions. If unstructured probing is required, the training may be longer, more complicated, and different in content than if structured questions are added to the questionnaire.

If a decision is made to use frame-of-reference probing questions for a subset of respondents rather than for all of them, additional interviewer instructions may be necessary.

Data analysis is the final step in the operation of frame-of-reference probing. Analysis focuses on responses to the probing questions and may also include their relationship to some of the other subjects of interest in the survey. Take, for instance, the example cited earlier in which respondents are asked to consider the seriousness of the following statement: "An offender injures a victim and the victim dies." Do people who imagine the injury to be inflicted during a barroom brawl rate the seriousness of the crime the same as or different from people who imagine it to have been the result of a mugging--or from those who imagine the death to have occurred as a result of a traffic accident? Differences in the responses of male versus female respondents or consistencies in the pattern of a single respondent's replies to a variety of such vignettes may also be of interest. If there is no differentiation among the rankings of crimes which are considered quite different by the questionnaire designer, there may be either a problem with the language in the question (suggesting that the wording should be changed), a problem with the researcher's notions about the seriousness of the crimes (suggesting that different examples be included), or perhaps a problem with the respondent's ability to make the desired distinctions (suggesting that the question should be deleted). Such an analysis conducted in conjunction with the final survey may provide explanations for some of the results from the analysis of the survey data.

E. Time Considerations

For the most part, the time required for planning and executing frame-of-reference probing overlaps preparation for the survey or test to which it is being appended. The selection of the testing vehicle, the data collection, and the data analysis all occur simultaneously with operations for the test or survey. Thus, the additional time necessary to use this technique is minimal. Drafting the probing questions (or deciding what information is required from unstructured probing) cannot take place until after the questions containing the words or phrases of interest are written, and it must be done before the interviewers who will administer the questions are trained.

Analysis of the information collected from unstructured frame-of-reference probing may take longer than from structured probing, since an additional coding phase may be required.

F. Cost Considerations

In general, the cost factors involved in frame-of-reference probing, over and above those of the test or survey itself, are slight. Additional expenses may be incurred for reproduction of questionnaires or interviewing materials, interviewer salaries for longer interviews, and salaries for the researchers/questionnaire designers. If members of the research staff conduct the interviews, cost of travel and related expenses, and extra salary expenses will also be incurred.

G. Mode of Data Collection

Frame-of-reference probing is suited for use in designing interviewer-administered surveys, either face-to-face or telephone. It could also be used in a face-to-face test of a mail questionnaire, but mail questionnaires themselves are not well-suited to the technique. Structured follow-up questions could be incorporated into a mail questionnaire, but since the respondent is free to answer questions in any order and over a long period of time, the responses to the probing questions may not be good indicators of what respondents had in mind when answering certain questions.

III. EXAMPLE: PILOT CITIES VICTIMIZATION SURVEY

The Pilot Cities Victimization Survey was conducted in 1971 to develop the National Crime Survey (NCS). It was a household survey in which respondents were asked the number and type of crimes committed against them and some details about each crime; in a portion of the sampled households, attitude questions about selected topics related to crime were included. Development of the questionnaire used in this survey and other work to develop the NCS was quite extensive and used a variety of the techniques described in this report. (See Example 1 in Chapter 10 for a description of another segment of the testing for this survey.)

For the purpose of this example, refinement of only the attitude questions will be discussed.

Two of the questions proposed for the survey were as follows:

> "Within the past year or two, do you think crime in your neighborhood has increased, decreased, or remained about the same?"
>
> "Would you say in general that your local policemen are doing a good job, an average job, or a poor job?"

If the study of attitudes about "neighborhood" was to be meaningful, some understanding of how large an area the respondent had in mind was required. In addition, unless information was obtained about what people considered to be crimes when they answered the question, researchers would not know what was viewed as having increased, decreased, or remained the same. Similarly, to interpret answers to the question about quality of police work, one would have to know something about what qualified as "good" and what qualified as "poor."

For these subjects, questions were prepared in advance and printed in a supplemental booklet (separate from the main survey questionnaire). Since the subjects were not considered to be particularly sensitive nor likely to bias the remainder of the survey questions, the questions to probe the frame of reference were inserted into the questionnaire immediately after the questions under study--that is, after each of the two questions cited above, the interviewer was instructed to go to the supplemental questionnaire, ask the appropriate questions, and return to the main questionnaire.

Concerning neighborhood, respondents were asked to describe the size of the area they considered to be their neighborhood; they could answer in terms of the number of blocks or miles, or could give names of streets and roads that bounded the area. In addition, respondents were asked whether they thought specifically about these boundaries in answering the previous survey questions.

To determine what "good" and "poor" police behavior was to each survey respondent, a list of 12 "typical" police behaviors was developed (e.g., enforcing all laws, shooting a looter who tries to escape, chasing away people who hang around streets or in doorways). After each item was read to them, respondents were asked whether they thought it represented "good" or "poor" police behavior. In addition, respondents' thoughts when the original survey question was asked were solicited (e.g., "Were you thinking about the actions of a particular policeman?" or "Were you thinking about something that happened to you?").

A similar exercise was used to probe the respondent's interpretation of the term "crime." Two of the questions used were--

> "Speaking of crime, everyone agrees some acts are crimes, but there are different ideas about others. Do you believe it is a crime for someone to ..."
>
> hold up a person?
>
> beat your wife?
>
> pass a bad check?
>
> sell liquor?
>
> litter the street?
>
> etc. (eight more acts were listed)
>
> "What kinds of acts were you thinking about when you said crime in your neighborhood is (increasing/decreasing/remaining about the same)?"

Since the questions were preprinted, recording responses was done easily on the supplemental questionnaire. While the questions were intended to add meaning to the answers given to the survey questions, they could not serve as consistency checks on the survey questions. Therefore, no way to reconcile inconsistent answers was needed.

About 80 interviews were administered during this phase; members of the research staff conducted all of the interviews.

Respondents for this phase of questionnaire development were not selected as part of a statistical sample; they were chosen because their house or apartment was in a census tract which had been selected for use in the Pilot Cities Victimization Survey.

Findings confirmed the suspicion that "neighborhood" was defined quite differently, even by next-door neighbors; therefore, the frame of reference for the question showed considerable variability. In this case, rewording to give a more precise reference of location was recommended:

"How safe do you feel on the street in front of your house?"

If a broader geographical area had to be included, then a question like the following could be tried:

"Would you feel safe in the streets anywhere in this city?"

For some respondents, "lots of policemen on patrol after 10 p.m." was "good"; for others it was wasteful and a sign of unwanted intervention in people's lives, and therefore, rated "poor." On many other topics, what was good police behavior to some was poor to others. Similarly, there was disagreement among respondents on whether some things (like selling liquor and littering the streets) were crimes. At best, the survey question could serve as a public opinion poll, but not as a measure of what type of police behavior satisfied people nor what people meant when answering the question about whether crime was increasing or decreasing.

Chapter 8
Observation and Monitoring of Interviews

I. INTRODUCTION

Observation of face-to-face interviews or monitoring of telephone interviews is most frequently thought of as a quality control technique, that is, a means of measuring interviewer performance and interviewer variability. This chapter examines the usefulness of observation and monitoring for a different purpose, that of evaluating the questionnaire and related data collection procedures. The term "observation" is commonly used in conjunction with face-to-face interviewing and "monitoring" with telephone interviewing, although both activities involve making similar sorts of judgments. In this report, "observation" is generally used in connection with both modes of data collection unless specifically stated otherwise.

Of the methods available to survey researchers for testing the adequacy of a questionnaire, observation of interviews is one of the most easily employed. Observation or monitoring to detect problems in the survey instrument and field procedures is conducted most frequently during the testing phase of the survey, including informal tests and formal tests. Clearly, this is the time when observational feedback can be of the greatest value in making revisions. However, a program of observation can provide researchers or survey designers with useful insights at any stage of data collection. For example, observations made throughout the interviewing stage of a one-time survey with an experimental or methodological component can be enormously valuable when discussing the results. Also, observations made during repetitive or continuous surveys can result in improvement in subsequent interviews.

Perhaps because the technique appears to be so simple, nonparticipant observation is rarely mentioned in the standard survey planning texts. Authors may assume that all survey designers routinely observe their questionnaires in action, although this is not the case. Commonly, observation or monitoring of interviews is considered the responsibility of the field supervisors rather than of the survey planners. Undoubtedly, this stems from the fact that interviews are usually observed to evaluate interviewer performance instead of questionnaire performance. Another reason why a discussion of observation and monitoring programs is usually absent in survey texts may be the seemingly subjective nature of the technique. The subjective element of

Principal Contributor: Deborah H. Bercini

a nonparticipant's observations allows for an unconstrained overview of the questionnaire and interviewing situation that is conducive to creative diagnosis of problems and formulation of solutions. However, the degree of subjectivity and reliability of observation is highly dependent on the system used to record the observations. Later in this chapter, various methods for recording interviewer behavior and questionnaire performance, some of which are rigidly standardized, will be presented.

Observation of face-to-face interviews or monitoring of telephone interviews by a third party who has been involved in the design of the survey, questionnaire, or data analysis plan can identify flaws in the data collection instrument and other procedures that cannot be detected by statistical analysis of the data or feedback from interviewers alone. Interviewers, no matter how skillful, are too involved in eliciting a response to "step back" from the interaction and fully analyze difficulties in communication with the respondent. As pointed out in Chapter 5 on informal tests, experienced interviewers may inadvertently conceal a defect in the questionnaire design by their ability to handle awkward situations. On the other hand, less experienced interviewers may attribute problems to the instrument that are more related to poor interviewing technique. Interviewer debriefings and written evaluations are extremely useful tools for judging the adequacy of a questionnaire. (See Chapter 9 for a description of the procedures and objectives of these techniques.) However, they cannot substitute for the observations of someone who is thoroughly familiar with the concepts and objectives of each questionnaire item.

The following is a compilation of some of the interview characteristics and questionnaire design issues that lend themselves to evaluation through observation or monitoring. The list is presented in a field test context, although many of the same characteristics can also be studied during subsequent stages of the survey.

A. Respondent Cooperation

Among respondents who agree to be interviewed, degrees of cooperation can vary greatly. The standardized explanation of the purpose of the survey and the confidentiality statement (if appropriate) that precedes the first question or a new series of questions must both motivate and inform respondents. An observer can note whether respondents understand the task they are being asked to perform by the questions they ask the interviewer or by irrelevant responses. The willingness of respondents to search their memory for requested information can be ascertained by the quickness or off-handedness of responses. If the consensus among observers is that respondents are reluctant to put forth the effort necessary to provide complete, accurate, or "valid" responses, then the survey instrument becomes suspect.

B. Interview Flow

A nonparticipant observer is in a particularly good position to judge whether the interview flows smoothly, and if not, to analyze the causes. Respondent confusion, distraction, or dwindling interest can be related to abrupt transitions between questionnaire topics or awkward and lengthy gaps, for example. Interviewers may have difficulty managing poorly formatted questionnaires,

or multiple questionnaire booklets, whether the interview is conducted face-to-face or over the telephone. The physical appearance of the questionnaire can encourage or frighten respondents, and observers can easily make note of this. A third party can also check whether flashcards or other materials handed back and forth between respondent and interviewer are aids or impediments to the progress of the interview.

C. Length of Interview

Interviewers are routinely instructed to record the beginning and ending times of an interview, so the overall length is almost always available. But nonparticipants can unobtrusively time individual sections of the interview and note the occurrence of substantial interruptions. Observers can make notes relating the time to characteristics of the household or the person being interviewed, such as the number of household members, health of the respondent, or other factors relevant to the survey topic. Because an observer does not have to be concerned with recording the responses, (s)he can be alert to cues that the respondent is losing patience, becoming fatigued, etc. The respondent's perception of the amount of time the interview is taking as manifested by comments such as "How many more questions are you going to ask?" is as valuable a piece of information as the measured interview time.

II. METHOD

A. Personnel and Skill Requirements

For the most part, the personnel involved in the observation of interviews for questionnaire design purposes are members of the survey staff who have been involved in planning the survey design, questionnaire, data analysis, or interviewer training. It is important to ensure that people familiar with all aspects of the subject matter, objectives and procedures of the survey provide advice during the development process.

Depending on the type of system used to record the results of observations, one or more coders may also be required to tabulate and summarize the results.

B. Selecting the Interviews To Be Observed

The primary purpose of a program of observation is to detect questionnaire and interviewing problems based on use in situations similar to those expected in the actual survey. Since this is also the general objective of a field test, formal or informal, the composition of the test sample is usually appropriate for a program of observation also. However, it is frequently not possible (and perhaps not desirable) to observe every interview in a field test. The survey researcher then must decide whether the kinds of observational feedback needed will be obtained from observations of a self-weighting, "representative" subsample or from observations of a biased subsample that includes a disproportionate number of units likely to provide a test of selected sections of the questionnaire.

For telephone surveys, the method used to identify a sample of interviews to be monitored depends on the sampling frame of the survey itself. The

selection of interviews to be monitored in a random digit dialed telephone survey field test cannot be as controlled as for a field test of personal interviews, because nothing is known about the sample unit before it is contacted. (In random digit dialed telephone surveys, the sample telephone numbers are generated randomly by computer.) Monitors should be aware that a large proportion of numbers dialed will be nonhousehold numbers, no-answers, or other forms of noncontacts. If the test sample for a telephone survey is in the form of a list of numbers known to contain eligible sampling units, then the selection of interviews to be monitored can be much more efficient.

Besides observing "live" interviews, another option available to survey planners involves tape recording the interview for detailed analysis afterwards. Respondent permission is necessary when this is done.

For all programs of observation or monitoring, it is particularly important that a variety of interviewers be selected so that observations are not biased by an interviewer effect. When monitoring telephone interviews, the monitoring schedule should cover as many interviewers as possible at different times of the day. For the same reason, it can be helpful to get feedback from as many observers as possible.

C. Preparation

1. Characteristics of Individual Questionnaire Items

To evaluate questionnaire items, an observer must have some notion of what constitutes acceptable question performance. Most researchers or survey planners probably feel that they will be able to detect question flaws without establishing a strict set of mental or written criteria. However, it is useful to learn what researchers in the field of questionnaire evaluation through observation have determined to be characteristics of successful questions.

Cannell and Robison (1971) set forth two basic dimensions for judging the adequacy of a question: How well the question communicates with the respondent, and the extent to which the question builds and maintains the relationship with the respondent.

Morton-Williams (1979) in an elaboration of Cannell and Robison's work, developed nine criteria for judging question performance.

a. The interviewer should have no difficulty asking the question correctly.

b. The interviewer should have no difficulty determining whether the question should be asked.

c. The question should be unambiguous.

d. The question should be about a subject that has meaning and relevance for the respondent.

e. The question should ask for information that the respondent is able to remember or has easy access to.

f. The question should ask for information that the respondent is willing to give.

g. The type of answer that is required from the respondent should be clearly conveyed by the wording or format of the question.

h. The objectives of the question should be clear so that the interviewer can decide if the responses should be probed.

i. The format of the question should make it easy for the interviewer to record the answer correctly.

On the assumption that a well-designed question will cause few problems for the interviewer or the respondent, survey researchers often evaluate questions by some of the same criteria that are used to evaluate interviewer performance. For example, individual questions are judged by whether the interviewer asked the question exactly as worded, asked the question in the correct sequence, omitted the question in error; whether the respondent asked for clarification, gave an adequate response, and so on.

In addition to general criteria which can be applied to almost any questionnaire item, observers usually evaluate the interviews against a set of very specific standards applicable to the individual questionnaire. For example, observers may note whether respondents consulted their bills and receipts for certain questions in a household expenditure survey or the ease with which the interviewer administers a complicated procedure that depends on the respondent's answer to a previous question.

2. System of Quantifying Observations and Training of Observers

For the observation/monitoring program to be of value to the questionnaire designer, the feedback from the observations must be relayed in a manageable, analyzable form. Similarly, the researcher or questionnaire designer must provide observers with some focus or objectives for their activities. Observers who are instructed to "note any problems" will probably return with a hodge-podge of unrelated comments that would be difficult to interpret. The survey planner must decide on the types of information (s)he wants to get out of the series of observations before the observations begin. The most useful feedback will come from observers who understand what specific problems and behaviors to look for and who have the ability to recognize the unanticipated rough spots as well.

The degree of structure imposed upon the observations will depend upon where the questionnaire is in its evolutionary development. The observational objectives for a questionnaire in an early draft form may be less defined because the survey planners are not yet fully aware of what the potential problems might be. As the questionnaire becomes more refined, so can the focus of the observations.

a. Using forms to quantify observations

Observations may be recorded on forms developed specifically for that purpose or observers can write comments directly on the questionnaire. If the survey planner wants to collect comparable information from each observer, it is advisable to use a standardized observer's form or observer's questionnaire. An observer's questionnaire can be constructed so that the observations are recorded in a standard fashion next to each questionnaire item. This is accomplished by inserting the observer's check item after each regular questionnaire item. Observation forms are often designed so that the same information is collected for each question, e.g., "question asked as worded," "question omitted in error," "respondent asked for clarification," and so on. Or the researcher may be interested in different but specific characteristics of some or all of the questionnaire items. In addition to the closed-ended, "check box" observations, more analytical, creative comments can also be gathered. In all cases, observers need to be trained on the use of the forms and the kinds of observations to record.

b. Verbal interaction coding

The kinds of observations that can be recorded during an interview are somewhat less detailed than those that can be obtained from analysis of a tape-recorded interview. Cannell et al. (1971, 1975) and Morton-Williams (1979) used tape-recorded interviews to develop and apply a coding scheme based on specific pieces of interviewer and respondent behavior, called verbal interactions. Each question was subjected to the same codes so that problem questions could be identified by the number and type of codes they received. Cannell's research (Marquis, 1971) involved the application of 52 specific behavior codes to 164 tape-recorded face-to-face interviews. Eight specially trained coding clerks coded the interviews. Agreement on which code to select was generally high (an inter-coder reliability of 86 percent was achieved) when coders could agree on whether a codable behavior had occurred. The following code categories, reduced from the original 52, were used in the analysis of the verbal interaction data.

Question Codes:

- Correctly asked question
- Incorrectly asked question
- Partial question
- Alternatives incomplete question
- Question omitted by mistake

Probe Codes:

- Repeat question
- Nondirective probe
- "Anything else" probe
- Directive probe
- Interviewer repeats answer

Clarification Codes:

Interviewer gives clarification
Respondent asks clarification

Response Codes:

Inadequate response
"Don't know" response
Refusal

For each question, the average number of problem codes were calculated, based on the number of times the question should have been asked. Thus, questions with code categories that had high average frequencies were considered inadequate in some respect. By grouping codes in various ways, the types of problems could be identified and attempts made to diagnose their nature. Three basic kinds of problems were identified--interviewer problems, respondent problems, and problems with the questions. The possible diagnoses included problems with question wording or context, problems due to lack of understanding of the underlying concept, problems indicated by erroneous omission or inclusion, and problems of refusal.

In evaluating his procedure, Cannell acknowledged that its usefulness would be enhanced by simplification. A major deficiency resulted from the fact that a single behavior can have many causes so that the technique could not always differentiate the nature of the questionnaire problems. But Cannell concluded that the procedure had "considerable potential for use in tests to locate problem questions and to provide adequate information which will permit the study director to correct the problem. The use of the procedures may make a substantial contribution toward objective evaluation of questionnaires at test stages."

Morton-Williams (1979) used a similar but somewhat more detailed verbal interaction coding frame to evaluate a questionnaire in its testing phase. She considered it a valuable, although expensive and time-consuming, technique. To achieve an acceptable level of reliability, coders had to be highly trained, not only in the application of the specific codes but also in proper interviewing technique. However, Morton-Williams recommended that questionnaire designers code a few taped test interviews because it would help them to think precisely about the objectives of each question, the task being asked of the interviewer and the respondent, and whether the question is appropriate and the instructions adequate.

D. Operation

1. Interviewer Training

The program of observation should begin at the interviewer training, even for informal tests. An observer/researcher who is confident that the interviewers received adequate preparation is in a better position to attribute difficulties in the interview to characteristics of the questionnaire or to the particular interview situation. If survey designers are made aware of

shortcomings in the training, they may be able to reserve judgment on certain troublesome sections of the questionnaire.

2. The Observation Setting

It is possible that the presence of an observer in the face-to-face interviewing situation will have an effect on the interviewer's and respondent's behavior, and thereby influence the data collected. These effects can be minimized, however, by a polite but brief introduction of the observer to the respondent and an unobtrusive manner of the observer. Usually the interviewer, after identifying herself/himself and gaining entry to the household or establishment, introduces the observer with a simple, factual statement such as, "This is _____ from (agency). He/she helps design the questionnaires we use." An advantage of using this introduction is that it gives the observer a legitimate reason to probe the respondent's answers at the end of the interview based on observations made during the interview. During the interview, observers should do as little as possible to remind either the interviewer or the respondent of their presence. If possible, observers should sit so they are not in the direct line of vision of either of the interview participants. Page-turning and note-taking should be done inconspicuously, and the observer should not interrupt during the interview.

Interviewers need to be reassured that the purpose of the observation is not to judge their performance, but to see how the questionnaire affects their performance. In household interviews it is generally considered unwise to pair a male interviewer with a male observer since respondents are often reluctant to let two strange men into their homes. The topic of the interview might also make it advisable to send out observers (and interviewers) of a particular sex. Of course, when the interview is conducted by telephone or tape recorded, these restrictions do not apply.

When properly conducted, an observation program for face-to-face interviews need not interfere with interviewers' schedules or delay the normal progress of the field test. Monitoring of telephone interviews can be accomplished with virtually no disruption whatsoever.[1] Similarly, tape recording interviews requires no deviation from the usual interviewing routine. Of course, the interviewer must get the respondent to sign a consent form giving permission to tape record the interview. The tape recordings are subject to the

---

[1]Regulations concerning "listening-in" or monitoring Federal telecommunications activities appeared in the Federal Register, Vol. 46, No. 61, March 1981 (41 CFR Pat 101 - 37). It stated that "consensual listening in," in which at least one of the parties consents to the monitoring, is permitted for the purposes of "service monitoring," where the monitoring is needed to effectively perform the agency's mission. Federal agencies conducting telephone interviews in-house or under contract vary in their interpretation of the regulations. Some agencies do not require that respondents be informed or give their consent to monitoring by a third party. These agencies maintain that the consent of the interviewer classifies the listening in as "consensual" and that the monitoring is needed to effectively conduct a telephone survey. Other agencies inform respondents that monitoring may take place with a statement such as, "My supervisor may listen to this interview."

same protections of privacy and confidentiality as the completed questionnaires.

Any time the survey researcher spends with an interviewer, such as time spent driving from one address to another during an observation session, can be used as an informal debriefing in which the interviewer is encouraged to comment on the questionnaire.

3. Obtaining Observers' Reports

Besides collecting and analyzing observation forms and coding sheets (if they have been used), researchers can gather additional insights by requiring written reports from observers. An observer may find it impossible to note all of his or her thoughts during the course of the interview. By reviewing notes from several interviews and summarizing the information in a single report, the observer has an opportunity to develop ideas for improving the questionnaire. The sooner these reviews are written following the observed interviews, the more valuable detail they will contain. Written reports also provide the survey planner with a manageable and permanent record of results.

Another extremely useful method for collecting results of the observation program is the observer debriefing session, although a debriefing session in which the questionnaire and interviewing procedures are reviewed is not necessarily a replacement for the written report. (See Chapter 9 for a description of interviewer debriefing sessions; the procedures for conducting an observer debriefing session are similar.)

E. Time Considerations

For the most part, the planning and execution of a nonparticipant observation program occurs in conjunction with planning and carrying out a formal or informal questionnaire field test, and the time required for such activity does not add to the total time allotted for questionnaire development. In fact, observation is one of the subjective evaluation techniques which are of primary importance in informal tests (as mentioned in Chapter 5) and which are of secondary importance in formal tests (as mentioned in Chapter 6.) The tentative time schedules presented in those chapters for carrying out those tests include the time necessary to incorporate an observation program into the test.

The more important time constraints concern the amount of time devoted to such a program by the professional staff, rather than the amount of time it requires in a questionnaire development schedule. The survey planning staff must spend a considerable amount of their time in a nonparticipant observation program, whether they are observing face-to-face interviews, telephone interviews, or listening to tapes. There may be other demands on staff time which force choices about what types of activities can be managed--for example, if researchers conduct the interviews in an informal test, there may not be sufficient personnel available to observe interviews. During the evaluation phase of the test, preparing observers' reports, listening to tapes, and possibly preparing reports of monitoring may compete with the

time required for the evaluation of data collected through other subjective techniques such as interviewer debriefing (described in Chapter 9).

F. Cost Considerations

The largest cost factor in an observation program is professional staff salaries. Depending on the geographic location and dispersion of the sample being observed, travel costs and related expenses for the observers may also be considerable. Otherwise, nonparticipant observation is a relatively low-cost way to improve the quality of questionnaire drafts.

G. Mode of Data Collection

This technique is obviously suited for use with interviewer-administered questionnaires, either face-to-face or on the telephone. It cannot be applied in mail surveys.

III. EXAMPLE: FIELD TESTING THE NATIONAL HEALTH INTERVIEW SURVEY EVALUATION QUESTIONNAIRE

A. Introduction

The National Health Interview Survey (NHIS) is a repetitive survey which collects health and demographic information from a national sample of about 40,000 households each year. Field operations for the survey are performed by the Bureau of the Census under specifications established by the National Center for Health Statistics (NCHS).

With the objective of fielding a revised NHIS questionnaire in the early 1980's, a series of field tests was planned to test an evaluation version of the NHIS questionnaire. The evaluation version, or experimental questionnaire, was designed to eliminate redundancies, define health concepts more explicitly, present topics in a more logical order and enable interviewers to use the material efficiently and smoothly. In conjunction with the results of a statistical analysis of the test data, the feedback from an extensive program of observation provided the basis by which to judge whether the objectives of the redesign had been achieved.

The testing was conducted in three phases.

1. Phase I (June 1979). The first version of the evaluation questionnaire was administered in 260 households in Springfield, Ohio, by 15 Bureau of the Census interviewers. The primary purpose of this informal test was to form a qualitative or subjective assessment of the draft instrument.

2. Phase II (October-December 1979). This phase of the testing was designed as a formal (split sample) test to quantitatively assess the revised evaluation questionnaire by comparing selected estimates produced by the standard NHIS document and the experimental document. The control group, consisting of the fourth-quarter 1979 NHIS sample (10,500 households), received the standard questionnaire. The experimental group receiving the evaluation questionnaire contained

5,000 households selected in the same manner as the control sample. Randomization of questionnaire versions among interviewers was not possible because of the risk that the 1979 fourth-quarter estimates from the continuing survey (the control group) could be affected by interviewer confusion of the two complex sets of rules and procedures. Instead, a separate group of interviewers administered each questionnaire version, the groups being matched as closely as possible on years of experience with the NHIS. The interviewers that had to be hired to meet the 50 percent increase in overall sample size were equally distributed among control and experimental groups.

3. Phase III (August 1981). Based on the outcome of the Phase II experiment, the evaluation questionnaire was again revised and used in an informal test in York, Pa. Like Phase I, the purpose of the test was largely qualitative. The size of the sample and interviewing staff were also similar to those in Phase I.

B. Programs of Observation: Phases I and III

Since the design and objectives of the Phase I and III informal tests were similar, their observation programs can be described together. Because both NCHS and the Census Bureau are involved in conducting the NHIS, observers from both staffs took part in the tests. The NCHS observers represented all of the disciplines involved in the survey's development, including questionnaire design, data analysis, and methodological design. The Census Bureau sent field supervisors and persons responsible for writing the training material and the interviewers' manual. Such a large and diverse observation team allowed for broad coverage of interviewers and a range of professional experience by which the adequacy of the training and questionnaire could be judged.

The test site and sample of households were selected by Census Bureau specialists in accordance with demographic, budgetary, and other procedural requirements. The households to be observed were determined indirectly by pairing observers with interviewers so that all interviewers were observed for at least 1 day, but not more than 1 day, by the same observer. Observations were conducted throughout the 5-day field test period. Approximately half of the test interviews were observed. The interviewer training session and the interviewer debriefing were also observed.

Observers from NCHS relayed their impressions in three ways: (1) observation forms (see Figures I and II)--observers were asked to time major sections of the interview, pay particular attention to new or difficult questions and concepts and indicate whether questions were understood, needed elaboration, or were difficult to ask (some of these observations could be tallied to give an indication of how frequently each problem occurred); (2) observer debriefing--led by one of the questionnaire designers; (3) written reports--specifying problems and solutions.

Census Bureau observers attended a separate debriefing which focused on the training materials, training session, interviewers' manual, and questionnaire.

Figure 1. Observation Form for 1981 Evaluation Questionnaire

1979 PHASE 1 PRETEST: Springfield, Ohio

Observer:_____ Time Interview Began:_____:_____ Date:_____

Segment/Serial#:_____ Time Interview Ended:_____:_____ # of Persons in Family:_____

SECTION OF QUESTIONNAIRE	Beginning Time	QUESTIONS NEEDING PARTICULAR ATTENTION	COMMENTS
Inside Cover Booklet	:	(Wa/Wb boxes. Entry in LA box in C1.)	
Introduction Page	:	Limitation of Activity Intro. read? Yes [ ] No [ ]	
Limitation of Activity	:	(Does respondent understand question 7 and question 8?)	
Restricted Activity	:	(When 2+ conditions in 6a, does respondent have problem with question 7?)	
Ambulatory Care Page	:	Introduction read? Yes [ ] No [ ]	
2-Week ACV Page	:		
Health Status Page	:	Condition List Introduction read? Yes [ ] No [ ]	
Condition Lists	:	(reporting of previously reported conditions)	
Hospital Page	:		
Condition Pages	:	(Is accident probe in 4c asked when needed?)	
Health Indicator Page	:		
Demographic Background Page	:		

Figure 2. Observation Form for 1982 Final Pretest

1982 NHIS FINAL PRETEST - YORK, PENNSYLVANIA, AUGUST 1981

Observer _____ Segment/Serial_____ Date 8/ /81 No. of Persons in Family _____

Section	Special Instructions	Beginning Time*	Questions Understood	Respondent Needs Further Explanation	Question Delivery Difficult	Comments
HH Page (p. 1)	Focus on Q 9 Tenure Q 11 - Classification of Living Quarters, Q 15 -	___:___	[ ] Y [ ] N↙ Q ___ Q ___	[ ] Y↙ [ ] N Q ___ Q ___	[ ] Y↙ [ ] N Q ___ Q ___	
Household Composition (p. 2)	Note that the concept of "Reference Person" is new.	___:___	[ ] Y [ ] N↙ Q ___ Q ___	[ ] Y↙ [ ] N Q ___ Q ___	[ ] Y↙ [ ] N Q ___ Q ___	
Limitation of Activity (p. 4)	Skip instruction for Q's 4, 7, 11 Complex - note problems. Be sure persons reporting LA in Q 3a (housework) are also asked if limited in work (Q 5). Skip instructions for different age categories confusing. Be sure Q 14 is asked for right person.	___:___	[ ] Y [ ] N↙ Q ___ Q ___	[ ] Y↙ [ ] N Q ___ Q ___	[ ] Y↙ [ ] N Q ___ Q ___	
Restricted Activity (p. 8)		___:___	[ ] Y [ ] N↙ Q ___ Q ___	[ ] Y↙ [ ] N Q ___ Q ___	[ ] Y↙ [ ] N Q ___ Q ___	

*If an interruption occurs during the course of interviewing, please note in the section for comments.

NOTE: This figure shows the first page only of a 2-page form. The second page used the same format.

C. Program of Observation: Phase II

Organizing a program of observation for the national split sample test phase posed many more logistical difficulties than the single-site tests in Phases I and III, since interviews were spread out geographically and over time. Only the experimental group interviews using the evaluation questionnaire were observed.

1. Interview Observations. At least 1 interviewer in each of 12 regions of the country was observed. An effort was made to observe both experienced and inexperienced interviewers. About 12 to 16 interviews were observed for each interviewer.

   For each interview, observers completed a brief observation sheet. This form obtained times for the many questionnaire sections and provided space for comments. In addition, observers were given a detailed memo about potential problems in the questionnaire. It should be noted that all observers were extremely familiar with the data collection instrument and its underlying concepts and objectives. Based on their accumulated observations, observers were asked to submit a written report.

2. Interviewer Debriefing Sessions. After data collection had been completed, interviewer debriefing sessions were held in each regional office. These sessions were observed by NCHS staff and Census Bureau staff. Their written reports, summarizing interviewers' comments, were submitted to NCHS questionnaire designers.

3. Interviewer/Supervisor Evaluation Forms. Every interviewer and interviewer supervisor was asked to fill out a lengthy questionnaire evaluating the adequacy of the training materials, training session, interviewers' manual and the NHIS questionnaire.

4. Regional Supervisors' Debriefing Sessions. NCHS survey planners conducted and observed a debriefing session of the Census Bureau regional supervisors at the end of the data collection period. Because supervisors had conducted the interviewer training sessions and had observed all interviewers in their region, their comments on the adequacy of the training materials and questionnaire were valuable.

   In conjunction with the results of the quantitative data analysis which compared estimates of key health variables obtained from the two NHIS questionnaire versions, the results of the more subjective field observations led to important revisions in the experimental questionnaire. This version was then tested in Phase III.

D. Results of the Observation Program

The questionnaire currently used in the National Health Interview Survey is the product of this multistage test in which observational feedback was as important as statistical analysis of the data. The evolution of the questionnaire during the phases of testing is illustrated by the series of

questions asked to elicit reporting of visits to doctors during the 2-week period preceding the interview.

The NHIS concept of a doctor visit is defined as a consultation with a physician in person or by telephone for examination, diagnosis, treatment, or advice. This service may be rendered directly by the physician or by a nurse or other assistant acting under the physician's supervision or authority. The standard (1969-1979) NHIS questionnaire used three probes to elicit reporting of doctor visits. They were:

> "During the past 2 weeks, how many times did you see a medical doctor?" (Do not count doctors seen while a patient in the hospital.)[2]
>
> "During that 2-week period, did anyone in the family go to a doctor's office or clinic for shots, x-rays, tests or examinations?"
>
> "During that period, did anyone in the family get any medical advice from a doctor over the telephone?"

NCHS analysts suspected that the concept of physician visits was not being fully understood by respondents. Of particular concern was the underreporting of visits to certain types of medical specialists, such as ophthalmologists and psychiatrists. Also, visits in which the patient saw a physician's assistant rather than the physician, phone calls made to obtain prescriptions, advice or test results, and visits occurring in places other than the usual doctor-patient settings were overlooked by respondents.

1. Phase I Version

The first version of the experimental questionnaire was designed to communicate the comprehensive definition of physician visit to respondents. The new questions were worded as follows:

> "These next questions determine whether anyone has recently received health care from any kind of medical doctor--including general practitioners and any types of specialists, such as pediatricians, psychiatrists, ophthalmologists, and so forth. Also include health care received from a doctor's assistant or a nurse working under a medical doctor's supervision."
>
> 1. "During the 2-week period outlined in red on that calendar, how many times did -- see or talk to a medical doctor or assistant? (Do not count times while an overnight patient in a hospital.)"

---

[2]Parentheses around parts of a question indicate to the interviewer that the statement is to be included conditional upon circumstances reported earlier in the interview. In this case, the statement is read only if the individual has previously reported a hospitalization.

2. "(BESIDES THOSE TIMES) During that 2-week period, did anyone in the family see a doctor or assistant for any surgery or operations, shots, X-rays, medical tests or treatment, or physical or mental examinations? (Do not count times while an overnight patient in a hospital.)"

3. "(NOT COUNTING THE TIMES YOU HAVE ALREADY TOLD ME ABOUT) During the 2-week period, did anyone in the family receive health care at home or make any (other) visits to receive health care at a hospital, or doctor's office, a clinic of any kind, or any other place?"

4. "During that period, did anyone in the family get any (other) medical advice from a doctor or an assistant over the phone?"

Observers attending the informal Phase I test reported that the experimental questions were much too verbose. Respondents frequently interrupted the introduction to answer "No," and would then become irritated at being asked the remaining questions. Instead of communicating the scope of the doctor visit concept, the wordy definitions and qualifications seemed to badger the respondent.

2. Phase II Version

For the national split sample test, the introduction was shortened so that it became a transition statement between questionnaire sections while the function of defining the doctor visit concept was distributed among the follow-up probe questions. The probe about the nature of treatment received was eliminated entirely, while the types of telephone calls to be included were stated more explicitly. The questions were:

"These next questions are about health care anyone in the family may have recently received."

1. "During the past [the 2 weeks outlined in red on that calendar] how many times did -- see or talk to a medical doctor? [Include all types of medical specialists, such as dermatologists, psychiatrists, and ophthalmologists, as well as general practitioners.][3] (Do not count times while an overnight patient in a hospital.)"

2. "We are also interested in the number of times anyone received health care from a nurse or anyone else working with or for a medical doctor. (Besides the time(s) you just told me about) During those 2 weeks did anyone in the family receive care at home or go to a doctor's office, clinic, or hospital to receive health care?"

---

[3]Statements in brackets were read the first time the interviewer asked the question in the household.

3. "(Besides the time you already have told me about) During those 2 weeks did anyone in the family get any medical advice over the PHONE from a doctor, nurse, or anyone else working with or for a medical doctor? Include calls to get prescription or test results."

A comparison of the estimates yielded by the control group questionnaire and the experimental questionnaire showed that the experimental questionnaire produced the desired reporting patterns. Major changes in the questions were not deemed necessary; however, some awkwardness was noted during the field observations. Observers reported that the questions were still too wordy, that respondents often gave a negative response to the introduction and that respondents answered Question 3 before the instruction to "include calls to get prescriptions or test results."

3. Phase III Version

To remedy these deficiencies, further revisions were made in Questions 2 and 3 for Phase III, informal test. Question 1 remained unchanged.

2. "(Besides the time(s) you just told me about) During those 2 weeks, did anyone in the family receive care at home or go to a doctor's office, clinic, hospital or some other place to receive health care? Include care from a nurse or anyone working with or for a medical doctor."

3. "(Besides the time(s) you already told me about) During those 2 weeks, did anyone in the family get any medical advice, prescriptions or test results over the PHONE from a doctor, nurse, or anyone working with or for a medical doctor?"

4. Final 1982 NHIS Version

Following the Phase III test, the experimental or "evaluation" questionnaire was revised for the last time before becoming the standard core NHIS instrument in 1982. Consensus among observers and interviewers was that the questions were still unnecessarily verbose. Although the basic structure and concepts were not changed, the final version of the questions reflects the effort to reduce them to their essential elements.

"These next questions are about health care received during the 2 weeks outlined in red on that calendar."

a. "During those 2 weeks, how many times did -- see or talk to a medical doctor? [Include all types of doctors, such as dermatologists, psychiatrists, and ophthalmologists, as well as general practitioners and osteopaths.] (Do not count times while an overnight patient in a hospital.)"

b. "(Besides the time(s) you just told me about) During those 2 weeks, did anyone in the family receive health care at home or go to a doctor's office, clinic, hospital or some other place? Include care from a nurse or anyone working with or for a medical doctor. Do not count times while an overnight patient in a hospital."

c. "(Besides the time(s) you already told me about) During those 2 weeks, did anyone in the family get any medical advice, prescriptions or test results over the PHONE from a doctor, nurse, or anyone working with or for a medical doctor?"

In this example, the repeated qualitative assessments made by observers (and interviewers) resulted in a more efficient series of questions. Statistical analysis of the formal test data in conjunction with observers' evaluations indicated at what point the benefits of a thoroughly defined concept were outweighed by the costs of a verbose questionnaire.

Chapter 9
Learning From Interviewers

Interviewers are a key and often underrated element in the practice of survey research. They constitute the link between respondents and researchers, and in their direct contact with respondents, they can pick up valuable information which may be of interest to questionnaire designers. Although much has been written on the subject of interviewing,[1] the systematic exploration of an interviewer's knowledge has been seriously neglected in the literature.[2]

Two techniques may be employed to elicit information accumulated by interviewers during the course of their duties. Data can be obtained from interviewers either through discussions (referred to here as interviewer debriefings) or through written evaluations (referred to here as structured post-interview evaluations). These two techniques can also be combined--participants in group discussion sessions may be instructed to fill out questionnaires before or during the session.

In this chapter, each of these techniques is discussed. Examples of the use of both techniques are presented at the end of the chapter (rather than after the description of each one), and the kinds of information obtained by them are compared.

I. INTERVIEWER DEBRIEFING

A. Introduction

The term "interviewer debriefing" refers to the technique of verbal information exchange between the interviewing staff and the operations staff. Both

[1]Writings on the subject generally fall into one of two categories: the task of interviewing (e.g., Kahn and Cannell, 1957; Richardson et al., 1965; Moser and Kalton, 1972; Babbie, 1973) and research related to interviewer effects on survey responses (e.g., Hyman et al., 1954; Henson et al., 1973).

[2]For an exception, see Converse and Schuman (1974), which relates the thoughts of graduate student interviewers about their interviewing experiences. Pages 64-72 are particularly relevant to questionnaire design, although the insights contained there were obtained through written narratives rather than either of the techniques described in this chapter.

Principal Contributor: Theresa J. DeMaio

of these terms are used loosely. The interviewing staff can be comprised of researchers if they happen to be doing the interviewing for an informal test, and the operations staff can encompass either field operations personnel or research personnel as the situation warrants.

Debriefing can be conducted at various points in the life of a survey, from the first stages of informal testing to the final, large-scale data collection effort. At any or all of these stages, interviewer feedback concerning problems in the structure or wording of a questionnaire can be crucial to improving the survey data. During a field test or pilot debriefing, results may be useful in revising question wording and response categories, identifying sensitive questions, improving the flow of the questionnaire, and estimating the respondents' ability to answer survey questions. At the end of a survey, suggestions from the interviewers might be used to evaluate the performance of the questionnaire, to contribute to the analysis of the results, or to recommend future changes in repetitive surveys.

The results of the debriefing process are qualitative rather than quantitative in nature. Although it can detect problems in the questionnaire (perhaps isolated among a particular population subgroup), the extent of those problems cannot be specified. While this may be seen as a disadvantage from a statistical point of view, the compensating advantage is that problems which were not anticipated by the survey designers (and thus not included on a form intended for statistical tabulation) may also be uncovered.

Interviewer debriefing can take two forms: group sessions or individual exchanges. Group debriefing sessions are a specific type of qualitative group interview (described in detail in Chapter 2), and generally consist of group meetings of survey interviewers, with a field supervisor or project staff person leading the discussion. Individual debriefings involve one-to-one communication between an interviewer and a supervisor, either in person or on the telephone. Group debriefing sessions occur more frequently and are discussed more fully here than individual debriefings.

B. Method

1. Personnel and Skill Requirements

A critical participant in the group debriefing session is the discussion leader. Several qualities are desirable in a discussion leader, although it may not be possible to find them all in one person. First, someone involved in the development of the survey will be familiar with issues that were problematic in designing the questionnaire and may note comments that might not seem important from some other perspective. While this has definite advantages, it also has some disadvantages. A discussion leader who has been intimately involved in a survey's development must not be defensive when negative comments are made, as this could discourage interviewers from making constructive contributions. Also, the discussion leader must not lead the interviewers into confirming his or her own preconceived notions about the questionnaire. Second, a discussion leader who is known to the interviewers may promote a more active exchange if this makes the interviewers feel less inhibited in expressing their opinions. This, too, has its drawbacks: field supervisors who are responsible for maintaining standards of

productivity and who constantly remind interviewers to read questions exactly as worded may not be the best people, from a research point of view, to lead a discussion of ways in which questions were asked or reasons why interviewers were not able to get accurate responses to survey questions. Third, an experienced debriefing leader should be able to keep the discussion focused on relevant subjects instead of having it drift onto extraneous issues. And fourth, experience and skill are required to obtain participation from timid as well as aggressive interviewers.

When multiple debriefing sessions are held simultaneously, more than one person must obviously be available to serve as a discussion leader. If sufficient numbers of researchers and/or field staff are not available, another alternative is possible. Representatives of the survey designers or the survey sponsor (i.e., the organization or agency that requests the survey and provides the overall objectives and funding) may be experienced and knowledgeable as discussion leaders. They may also attend the sessions as observers of the proceedings, or as participants with a limited role in the discussion. This is particularly important if the discussion leader cannot view the discussion from the perspective of the survey objectives or the development of the questionnaire.

The degree to which the survey designers are involved in the debriefing process (as observers or discussion leaders) can determine the extent to which results are incorporated in questionnaire revision or analysis. A close working relationship between all parties involved in the process of improving the questionnaire is suggested for maximum results.

2. Selection of Interviewers

The number of participants in a debriefing session may vary according to the type of survey involved. In an informal test, the number of interviewers may be only five or six, while in a formal test or survey, the number might be much larger. Generally speaking, if the number of participants exceeds 15, separate groups should be assembled to allow for maximum participation by the interviewers. With smaller groups, more information can be obtained from each interviewer.

Depending on the geographic area encompassed by the survey and on constraints of budget and timing, it may be possible to hold debriefing sessions in more than one place. For example, in a national survey or a field test conducted in three areas of the country, two or three debriefing sessions might be arranged in different cities. Increasing the number and location of the sessions has two advantages: (1) it includes reports of experiences with respondents in more than one geographic region, who may have had different reactions to or problems with the questionnaire, and (2) it decreases the possibility that the results (of a single session) may be idiosyncratic due to particular interviewers' skills, supervisors' instructions, or discussion leaders' ability to control the group.

3. Preparation

In planning debriefing sessions, several elements need to be considered.

a. When to hold sessions

Successful results may be obtained during a field test debriefing either by conducting a single discussion at the end of the test or by conducting discussions on an ongoing basis (e.g., daily). Holding sessions more frequently and implementing changes in the questionnaire throughout the testing period allows a number of versions of a question to be tested, if necessary.

Regardless of the stage of the survey at which the debriefing session is held, it should be conducted very shortly after the end of interviewing. This ensures that the experiences of the interviewers will be fresh on their minds and more accurately reported.

b. How long they should last

The length of a debriefing session depends on the amount of material to be covered. The average session might last two or three hours, but all-day debriefing sessions are not uncommon. Discussions scheduled for longer than a couple of hours should be interrupted by breaks.

c. Outline

To ensure that the discussion covers appropriate, prespecified topics and maintains a proper focus, an outline should be prepared in advance of the debriefing session to guide the discussion. The content of the outline can include some topics which are important from the perspective of questionnaire design and some which are not (e.g., discussion of administrative or survey operations procedures).

The outline should include those features of the questionnaire about which the designers are most anxious to receive feedback. If different versions of a questionnaire or sections of a questionnaire are being tested, the interviewers' judgment about which version worked best (and their reasons for arriving at that judgment) should be solicited. The extent to which respondents seemed to understand particular words or concepts, had the information or were willing to answer particular questions, viewed particular questions as sensitive, etc., might be included as topics for discussion.

It is generally helpful to provide interviewers with some idea of the topics to be covered during the debriefing session. This can be done either by circulating an agenda containing questions for discussion prior to the debriefing session, or by handing one out at the beginning of the session. This will give the interviewers time to think about the issues and to recall relevant experiences; this promotes more informed discussion during the session itself. It also lets the interviewers know that particular topics will be covered so they will be less likely to interject their views at inappropriate places in the discussion.

4. Operation

One of the positive features of group debriefing sessions (as mentioned in Chapter 2) is that the group atmosphere promotes interaction among the interviewers and stimulates them to react to the ideas of others, possibly

increasing their own insights and thus the value of the discussion. It is the responsibility of the discussion leader to emphasize the importance of interviewers' input, both positive and negative, and to set the tone of the discussion. All parts of the debriefing session will not be equally productive from the questionnaire designer's point of view. However, allowing interviewers to vent their frustrations about some topics that are beyond the questionnaire designer's control will be necessary at some points. Some so-called "wasted" time should be expected during a session.

Interviewer debriefing sessions are generally tape recorded. This practice is useful because (1) it enables a more accurate transcription of discussions that move too quickly for a scribe to record and (2) if the debriefing report is not prepared immediately, it prevents the results from being subject to memory decay.

There is a drawback to this practice, however. The transcription of the debriefing tape is a time-consuming process, often completed after such a long lapse of time that the usefulness of the results in questionnaire revision is diminished.

Even when a tape recorder is used, it is a good idea to have a designated note-taker and to rely on the tape recorder only to review particularly noteworthy parts of the discussion and sections that moved too quickly for accurate note-taking.

After all scheduled debriefing sessions have been held, a summary of the main results should be prepared. The summary should include implications for questionnaire revision if the interviewing is conducted as part of an informal test or formal test, and it should be prepared as quickly as possible. Often, when a questionnaire is revised after a test, the exact changes and the reasons for making those changes are not documented. This has two drawbacks: it prevents others from learning from the experience, and it prevents anyone from knowing whether the debriefing results are used.

5. Individual Debriefings

The second method for conducting interviewer debriefings involves individual meetings of each interviewer with his or her supervisor, which can be scheduled at regular intervals or at the end of an interviewing period. Problems experienced in the field with the questionnaire, procedures, or particular respondents are topics for discussion. These meetings can take place over the telephone or in the office, perhaps when an interviewer turns in completed work. An outline is useful in this type of debriefing, too, and debriefers should use the same outline in talking with each interviewer. This type of encounter is more valuable as a quality control or interviewer support mechanism than as a questionnaire design technique--it does not give the survey designer a reading of the prevalence of questionnaire design problems, unless the outline is extremely specific. Interviewers may have different priorities about the problems to bring up with their supervisors. If only one interviewer mentioned a problem with the respondents' understanding of a particular question, the problematic aspects of that question may be severely underestimated.

6. Time Considerations

As is the case with observation and monitoring of interviews, interviewer debriefing is one of the subjective techniques used in the evaluation of informal and formal questionnaire field tests. The planning and execution of one or a series of interviewer debriefing sessions generally occurs within the context of these tests, and the tentative time schedules presented in Chapters 5 and 6 include the time required for the debriefing to take place.

In comparison with observation/monitoring, interviewer debriefing involves a smaller investment of time by the professional staff, a greater investment of time by the interviewers, and approximately the same amount of time between the end of interviewing for the test and the completion of summary reports.

7. Cost Considerations

Interviewer debriefing can be a relatively low-cost tool for use in the questionnaire design process. For a small-scale informal test conducted near the agency headquarters, travel costs are minimal and the salaries of the personnel involved would be the primary cost factor. Depending on the geographic area included in the test or survey, however, the cost of an interviewer debriefing program may vary considerably. For a national field test or survey in which multiple debriefing sessions are held throughout the country, the cost of travel and related expenses for the debriefers (and interviewers, if they are not located near the debriefing site) may far outweigh the cost of salaries. In addition, other minor expenses such as renting a debriefing site may be incurred when debriefing sessions are not held near agency facilities.

8. Mode of Data Collection

As the name implies, interviewer debriefing is suited for the evaluation of interviewer-administered questionnaires--either face-to-face or telephone interview schedules.

II. STRUCTURED POST-INTERVIEW EVALUATION

A. Introduction

Structured post-interview evaluations are often referred to as "ratings" and involve administering questionnaires to interviewers after their participation in the survey has been completed. The attitudes and behavior of an interviewer can influence a respondent's answers. These evaluations contain questions about interviewers' attitudes and perceptions of their respondents, which may provide input concerning potential sources of bias. Do the interviewers feel inhibited in asking for respondents' income? Do they view respondents as cooperative during the interview? Do they think the respondents give accurate and honest answers to the survey questions? How do interviewers feel about the objectives and value of the survey? Factors such as these might influence both the quality of the data provided by the respondents (when they answer the questions) and how often responses to the questions are not obtained.

The ultimate objective of such evaluations is to obtain information about the attitudes and behaviors of the participants in the data collection process that may affect responses to survey questions. In some instances, the results of these evaluations can be used to improve a questionnaire draft; in others, they can be used to improve a future wave of a survey; in still others, they can be used to give the survey designers or data analysts information about the kinds of errors that may have been introduced during the data collection process. In this last use of post-interview evaluations, the results are more likely to be incorporated as revisions to the procedures for interviewer training or data collection than as revisions to the questionnaire.

B. Method

1. Personnel and Skill Requirements

The project director for a post-interview structured evaluation program should have enough familiarity with sources of interviewer bias to formulate hypotheses about interviewer effects in the survey (or test) being evaluated, develop a questionnaire that collects data to test those hypotheses, and evaluate the data that are collected. Additional staff may be required to work toward completion of these tasks.

Use of this technique involves a minisurvey of a sort, and requires interviewers to serve as respondents. Most often these surveys consist of self-administered questionnaires; if face-to-face or telephone interviews are used instead, additional personnel (i.e., other interviewers or supervisors) are needed to perform the interviewer function.

In practice, post-interview evaluations are generally treated independently of the original field test (or survey) and are often organized and conducted by different groups of people. This can lead to two problems: (1) lack of coordination between the groups involved in developing evaluation forms and acquiring data for analysis; and (2) lack of incorporation of research results that might improve the survey. These limitations of the method can be minimized by conscious effort and communication between the two groups.

2. Selecting the Interviewers

Selecting the interviewers (i.e., the respondents to the evaluation survey) is not an issue. The participation of all the interviewers who are involved in the survey (or test) is generally requested in the evaluation. Because the number of interviewers involved is relatively small to begin with, and because the responses of all types of interviewers are important to the results, it is imperative that interviewers take the evaluation seriously and that all interviewers participate.

3. Preparation

Decisions concerning the content of the evaluation questionnaire depend on the researcher's hypotheses about sources of bias. Several kinds of perceptions can be solicited from interviewers: questions can be asked about the interviewers themselves, about the survey instrument, or about the

respondents. When interviewers are questioned about the respondents, a decision must be made concerning the unit of analysis for the data. Interviewers can be asked to complete a separate evaluation for each interview in their assignments, or they can be instructed to make a judgment about their respondents as a whole. Using the first approach, there will be as many evaluations as there are respondents; the second method can be disaggregated only to the interviewer level.

The first method is more cumbersome in planning and execution, but its results are more precise. Using the second method, an interviewer might be influenced in making his or her aggregate ratings by situations that were particularly memorable (as either good or bad experiences) but not typical of the entire assignment. Also, different interviewers have different abilities to generalize, so their estimates of "some," "most," etc., of their respondents may not be comparable.

The description of procedures for obtaining structured evaluations thus far has centered on their use after data collection for the survey or test has been completed. In addition, such evaluations may be used in conjunction with interviewer debriefings (discussed in the first part of this chapter). During the debriefing session, interviewers can be instructed to complete a questionnaire containing specific questions (perhaps the same questions that are discussed in more detail during the session). In this way, responses to every question can be obtained for every interviewer, which may not be the case in the less structured debriefing session. Another advantage of this technique is that quantitative results are obtained, which can be tabulated to provide a more specific idea of the extent to which specific problems or behaviors are occurring.

4. Operation

Although the data for post-interview evaluation can be collected either by means of self-administered questionnaires or interviews (face-to-face or telephone), it is usually done with self-administered questionnaires. This is less expensive than other methods and more practical, particularly when the evaluation design calls for interviewers to rate each respondent separately.

The evaluation data are obtained during the data collection phase of the test or survey being evaluated. If ratings of each respondent are required, an evaluation form should be completed at the end of each interview--before the interviewer approaches another respondent. If generalized respondent ratings are required, interviewers should complete a single evaluation form at the end of their interviewing assignments.

In most uses of post-interview evaluations, the collection of the evaluation data is part of a larger scheme. The next step in these evaluations is to link the data obtained from the interviewers with information collected in the survey or test. The importance of this technique in questionnaire design is to learn whether some aspect of the questionnaire, which can be changed, affects interviewers' attitudes. To determine whether the interviewers' perceptions had any effect on survey responses, some measure of the quality of those responses is necessary.

Two types of response quality indicators are available. One is, obviously, the data collected in the survey. The particular data items used to measure response quality can vary according to the hypotheses of the researchers. In general, investigators view interviewer ratings in relation to an indicator of data quality such as item nonresponse or level of reporting. Item nonresponse affects data quality because it affects the amount of imputation or the number of cases that can be used for a particular analysis. It also has the advantage of being easy to measure. Other indicators such as level of reporting require making an assumption about the relationship between that indicator and response quality--for example, the more doctors' visits or incidents of illness are reported, the better the data are assumed to be. This may be a reasonable assumption, but it is an assumption nonetheless. Better evidence of data quality (i.e., whether the questions were answered truthfully) may be very difficult to obtain. It involves obtaining independent corroboration of respondents' answers, either through record checks or evidence from another reliable source. (See Chapter 10 for a discussion of record checks.) This is not always possible, and even when it is possible, it may be quite expensive.

The second type of response quality measure is not directly related to data collected in the survey. Instead, an assumption is made that items contained in the evaluations are indicators of the quality of the data collected in the survey. For example, in collecting data for a consumer expenditure survey, an evaluation of the respondent's ability to provide information about expenditures may be assumed to reflect how well the expenditures were reported. Then, the items included in the evaluation questionnaire can be used as the dependent variables in the analysis. Care should be taken in this type of analysis to assure that the assumptions are reasonable ones.

When all is said and done, sometimes the results of this type of research are difficult to apply directly to the operation of a survey. For instance, even if research documents that interviewers with certain types of attitudes have lower response rates or item response rates, ways to alter those attitudes may not be obvious. Creative solutions to the answers, obtained by creative research, are also a necessary part of the process.

5. Time Considerations

Post-interview structured evaluations are more time-consuming than the other methods of evaluating a questionnaire discussed thus far. However, since most of the planning and analysis is ordinarily done by researchers rather than field personnel, its use need not add much time to the survey schedule. The data collection can be conducted simultaneously with data collection for the survey or test (if interviewers complete a form for each respondent) or at the very end (if interviewers complete only one form). The longest phase of the project involves data processing and analysis; the length of this phase depends on the stage of survey development at which it is used and the sample size (which determines whether the data are tallied by hand or by computer). In general, if the evaluation data are analyzed simultaneously with the survey or test data, the two tasks should be completed at approximately the same time.

6. Cost Considerations

The most expensive aspects of using this technique are professional staff salaries and data processing expenses for keying and analysis. The magnitude of these costs depends on how large the data set is (i.e., how long the evaluation questionnaire is, how many interviewers are involved, and how many evaluation forms are completed by each interviewer). The cost of analysis for a post-interview structured evaluation program is much less than comparable tasks for the survey or test itself.

7. Mode of Data Collection

A post-interview structured evaluation program is only appropriate for use with interviewer-administered questionnaires.

III. EXAMPLES

A. Example 1: Interviewer Debriefing on the Consumer Expenditure Survey

The 1972-73 Consumer Expenditure Survey was conducted by the Bureau of the Census for the Bureau of Labor Statistics to collect data used in constructing the cost of living index.[3] This survey also provided experience that was used to design a continuing consumer expenditure survey implemented in 1979. Interviewing for the 1972-73 survey used a long and extremely detailed questionnaire requesting information about types and amounts of expenditures in all categories of household expenses (e.g., mortgage payments and ownership costs, medical and health expenditures, house furnishings and related household items). The survey was structured to include five personal-visit interviews at each sampled household. Data for some types of expenditures were collected in each quarterly interview; other information was collected only in one or two quarters. After each interview, the interviewer told the respondent what types of expenditures would be included in the next interview, and a card or pamphlet to record them was left with the respondent.

This example was chosen for this report because it illustrates the use of post-interview evaluations together with interviewer debriefings and elaborates on the differences between the information obtained by the two methods.

At the end of the first year of interviewing for the survey, three debriefing sessions were arranged in various sections of the country. Twenty-one interviewers, most of whom had worked in all five interviewing periods, were assembled and their permission to have the meetings tape recorded was obtained. Discussions were led by members of one of the Bureau's research divisions, who were specialists in questionnaire design. Two staff members conducted the debriefing sessions; one led the discussion and the other served as an assistant.

The discussion of the questionnaire proceeded according to the following outline:

---

[3] Information presented here is adapted from Rothwell (1974a).

1. For the next few hours, let's reverse the pattern--you talk and we will listen. I suppose many of you know each other, but for those who don't know everyone else and for me, let's start by going around the table. By way of introducing yourself, would you start by saying how many quarters you worked on the Consumer Expenditure Survey and about how many interviews you have conducted? (Go around the table for this.)

2. Now I'd like to make one more trip around the table and have each of you give a few minutes of honest advice to an imaginary friend who is taking a job as an interviewer on the Consumer Expenditure Survey. What advice would you give your friend? What problems would you warn her about?

3. Without looking at the questionnaire, were there any questions which you remember as having bothered people, angered or annoyed them?

4. In what terms did people recall their purchases--that is, in what ways did they remember what they bought? Was it by month of purchase, weather conditions at the time, nearness to a payday, or was it by family member or in some other way?

5. What kinds of questions did your respondents have the most trouble answering? What probes or reminders worked best?

6. Now look at the questionnaire and circle any questions for which you think the information may not be very accurate or precise. After you do that, put a checkmark alongside of any questions that irritated people.

7. There were two cards like these[4] which were used in the survey. How did the white one which you left behind after the first quarter work for you? How many people used it? How many people used the yellow card? Would you recommend greater use of cards like these?

8. About how many of your respondents would you say kept budgets or kept track of how they spent their money? About how many referred to some records to answer questions?

9. Do you think it would be possible to get more respondents to check their records? How would you feel if you were instructed to ask them to do that?

10. Did you have any respondents who tried to get out of the interview once you started or who didn't want to cooperate after one or two interviews? About how many? What did you do to convince these people to continue?

---

[4]The cards referred to here informed the respondents about topics for the next interview and enabled them to record their expenses.

During the session (at item 6 on the outline), interviewers also completed a post-interview structured evaluation form requesting information about the section and item numbers that caused difficulty for respondents. (A portion of this form is included as Figure 1.)

The discussions lasted approximately 4 hours. Later, the tape recordings were summarized independently by two researchers, and differences of interpretation were reconciled. After agreement about the content was reached, the summaries were coded for the types of problems identified by the interviewers.

Analysis of both the debriefing summaries and the forms filled by interviewers were included in the final report. Information obtained from the discussions differed from the written comments in several ways that complemented each other:

1. The first way in which the content of the two methods differed was in completeness of coverage. Comments concerning some specific items or sections of the questionnaire were included on the forms but not mentioned during the discussion. This could have been a function of time pressures during the discussion, unassertive interviewers, abrupt shifts in the discussion, or lack of recall of the problem at the appropriate time to mention it.

2. There were several ways in which the two methods of input from interviewers yielded different types of information.

   a. The discussion pointed out general areas of difficulty for respondents (e.g., the respondents had trouble understanding the vocabulary in the section on home ownership costs, the concept of "consumer unit" in others). In contrast, the written comments provided specific item numbers that illustrated the problems.

   b. The discussion uncovered problems that the interviewers perceived as affecting data quality, but that did not result from respondent difficulty with a particular question, and thus were not included in any written comments (e.g., interviewer perceptions of under-reporting on certain types of expenditures based on their observations at the household).

   c. The discussion mentioned reasons that might account for respondent difficulty with particular questions, as well as possible solutions to some of the problems. This information was not obtained in the written comments.

   d. Written comments provided a more accurate estimate of the number of interviewers who had problems with questions, and a more exact enumeration of the questions that caused problems than did the discussion.

Thus, use of both methods in tandem provided more useful information than either method used separately would have obtained.

Figure 1. Example of a Structured Post-Interview Evaluation Form Used During Interviewer Debriefing

1972 CONSUMER EXPENDITURE SURVEY

Please circle the number of any sections that caused problems in general or contained individual questions with which there was difficulty. Next to the appropriate section record these individual question or item numbers.

Section	Number of Question/Item Which Caused Problems
1. Household Record and Consumer Unit Determination	
2. Rented Living Quarters	
3. Owned Living Quarters and Other Owned Real Estate	
4. Mortgage Payments and Ownership Costs	
5. Expenditures for Repairs, Alterations, and Maintenance of Owned Property	
. . .	
25. General Housing and Consumer Unit Information	
26. Work Experience, Income in 1972, and Other Selected Items	
27. Assets and Liability Changes in 1972	
Additional Comments:	

Substantive contributions of the interviewers, relevant to questionnaire design, fell into the following general categories:

1. Question wording. For example, interviewers suggested replacing "vehicle registration tags" with "license plates"; they also suggested that phrases be added to some questions to provide examples and clarify the intent of the question--"did you pay any refundable deposits for this unit, such as a security deposit?"

2. Question sequencing. For example, interviewers suggested combining questions on the same topic that were asked in different interviews--in later interviews, some respondents reviewed their records and felt trapped or embarrassed when they discovered they had inadvertently answered a question incorrectly in a previous interview.

3. Reference periods. For example, for certain types of items, interviewers felt that the reference period was too long; in other cases, the shift in reference periods was confusing to respondents.

4. Format and physical features of the questionnaire. For example, the cumbersome questionnaire contained many very large (11" x 16 1/2") pages, attached with wire spiral loops across the top. Suggestions were made to increase the size of the loops to facilitate turning the pages, and to print all the pages in the same direction to make the administration of the interview more convenient.

B. Example 2: Structured Post-Interview Evaluations on the Consumer Expenditure Survey

During the second year of interviewing for the Consumer Expenditure Survey (i.e., in 1973), a different type of research effort was conducted. Post-interview evaluations were obtained from both respondents and interviewers who participated in the survey and used to compare indicators of response quality from both participants in the data collection process. It was felt that in obtaining evaluations from both interviewers and respondents, more reliance could be placed on conclusions for which the ratings agreed than on those for which they disagreed.[5]

Follow-up evaluations were conducted after the final interviews at each household had been completed. To obtain respondent reactions to the survey, a 5-percent subsample of respondents was selected and interviews were conducted, principally by telephone, by supervisors or office staff performing a quality control function. The interviews averaged 10 minutes and included a series of questions about respondents' interest in the survey, the kinds of records they kept, their use of such records to answer the survey questions, and which topics included in the survey gave them difficulty. A total of 531 respondent reinterviews were completed.

---

[5]Material presented here is summarized from reports by Rothwell (1974b) and Glynn (1978).

As a separate activity, interviewers were asked to complete an evaluation form containing similar types of questions for each respondent in their fifth-interview assignments. This form was completed for nearly 99 percent of the respondents (10,122 cases) in the 1973 Consumer Expenditure Survey panel.

Independence of interviewer and respondent assessments was assured by the way the information was obtained. Since every interviewer was requested to fill a form for each respondent, interviewer assessments were not dependent on the ability to generalize about the reactions of all respondents. Also, interviewers had no idea which of their respondents would be selected for the reinterview sample or what questions would be included in reinterview. Therefore, they had no opportunity to make their own evaluations consistent with those of their respondents.

Although the independence of the assessments was assured, one flaw in the planning should be noted here. Since one of the aims of the research was to compare respondent and interviewer evaluations of the same attitudes and behaviors, the ideal situation would have been for the follow-up questionnaires administered to both groups to have contained the same questions, or at least the same response categories. However, due to a lack of coordination in the development of the forms, the two questionnaires used in this research are similar rather than identical.

After the data were collected, the questionnaires were coded and keyed. In addition, data for the respondent reinterviews were matched with the data for those respondents provided by the interviewers. Of the 531 completed respondent reinterviews, 506 were matched with the corresponding interviewer evaluations and constitute the base for the analysis included in the final report.

The analytic strategy was to evaluate interviewer and respondent reports against each other. The analysis conducted to achieve this goal was based on the belief that use of records--such as check stubs, receipts of purchases, or bills--to answer the detailed expenditure-related questions in the interview was an indicator of response quality. Respondents who checked their records before answering questions were assumed to provide better quality data than those who did not. Use of records, therefore, was the major dependent variable; the level of record use was measured through both respondent and interviewer reports, and its association with demographic characteristics was measured to investigate whether some types of respondents provide better quality data than others.

Analysis of the data revealed that four-fifths (80 percent) of the respondents reported that they kept some kind of records. Of those respondents who reported keeping some kind of records, 65 percent used them in answering questions about expenses; of the total sample, 54 percent of the respondents referred to records in replying to the Consumer Expenditure Survey. There were differences in the extent to which records were used to answer survey questions by members of demographic subgroups: older respondents seemed more likely than younger respondents to refer to records; females were more likely than males; whites were more likely than blacks (or members of other racial groups); respondents in households headed by individuals whose highest

level of education was 12th grade were more likely than respondents in households headed by individuals whose highest level of education was greater or less than 12th grade. In addition, there appeared to be little consistency between respondent self-reports of record usage and interviewer reports of respondent record use.

Respondents may not be able to answer questions for which they do not keep records or for which their records are poor. The 20 percent of the respondents who reported not keeping records were at a disadvantage in terms of being able to provide accurate answers to expenditure questions. Additional respondents who reported they kept records may not have used them because of their poor quality.[6] One implication of these findings is that questions about recordkeeping practices should precede questions about expenditures, and the expenditure items might vary according to the respondents' ability to answer them accurately.

C. Example 3: Structured Post-Interview Evaluation on the Telephone Health Interview Survey

A research effort was undertaken by the National Center for Health Statistics (NCHS) in 1978 to investigate the feasibility of conducting Federal health surveys using telephone rather than face-to-face interviews. This research (Bercini and Massey, 1979) was conducted in conjunction with a cigarette smoking supplement to the Telephone Health Interview Survey, and the indicators of data quality used were overall nonresponse rates and item nonresponse rates for the question requesting names of household members. It was selected for inclusion here because it illustrates the use of post-interview evaluations in conjunction with an experimental design, and because evaluations were obtained about the interviewers rather than about their respondents.

One difference between telephone and face-to-face interviews is that it is relatively easy for telephone respondents to discontinue the interview (i.e., hang up the phone) at any point, whereas once a face-to-face interviewer gets access to a house, it is less likely that the interview will be terminated.

The household roster (i.e., the section of the interview in which the household composition, names and demographic information about household members is obtained) is particularly subject to respondents' ending the conversation because of its sensitive nature and its seeming lack of relevance to the stated purpose of the interview. This section of the questionnaire, therefore, was a suitable subject for investigation concerning ways to reduce nonresponse. Accordingly, an experiment was designed to see (1) if obtaining the household roster at the end, rather than at the beginning, of the interview would affect response rates, and (2) if obtaining the household roster without asking for the names of the household members would affect response rates.

---

[6] There are other explanations for this finding as well, such as lack of motivation on the part of the respondent.

A 2 x 2 factorial design was employed and four versions of the questionnaire were developed. The four versions were randomly distributed to interviewers, who conducted interviews using more than one version. An alternative approach of randomly assigning interviewers to conduct interviews using only one questionnaire version was not feasible, although it would have had the advantage of controlling for the effects of interviewers' preferences for one version over the other.

After the interviewing, interviewers' evaluations were obtained. They were asked to rate the experimental questionnaires in order of preference and ease of administration. Self-ratings were also obtained of how reluctant they were to ask for names of household members and how persistent they were in obtaining names from hesitant respondents.

The data for analysis included evaluations from 19 interviewers and the outcomes of attempted interviews with initial respondents at 2,565 eligible households. Three different types of overall response rates and an item nonresponse rate for names of household members were calculated from the survey data.

- Placing the household roster at the end of the interview rather than at the beginning significantly improved the response rate.

- Asking for the names of household members appears, in some cases, to reduce response rates. However, this difference appears to be more attributable to the impact of interviewer behavior than to questionnaire differences.

- Interviewers' rankings of their preferences for the questionnaire versions generally coincided with the level of response rates which were obtained, suggesting that sensitive questions can have a significant impact on interviewer attitudes and performance.

- Interviewer reluctance to ask names is associated with lower response rates.

- Interviewer persistence is less clearly associated with response rates. Interviewers who are highly persistent have low response rates; those who are somewhat persistent, and who presumably know when enough is enough, are the most successful in terms of response rates.

The results of this research might be used to improve the design of telephone interview questionnaires. The implications of the interviewer evaluation findings, in particular, however, are applicable to interviewer training and selection.

Chapter 10
Using Record Checks

I. INTRODUCTION

The techniques discussed in Chapters 7, 8, and 9 have used the assessments of participants (i.e., interviewers, respondents, and observers) as evaluation criteria. These tools can provide information about some important aspects of questionnaire design, but they cannot necessarily determine whether survey questions are being answered accurately. This chapter describes the use of records and matching as an independent source of evaluation.

A match is any linkage of records from the same population to provide more complete information pertaining to an individual or a group. Matches are either exact or statistical; that is, the linkage brings together information from different sources for a specific person or it associates data for persons who have similar characteristics. Record checks, as used here, are a form of exact matching. In the development of questionnaires, the purpose is not so much to accumulate more information about an individual, but to compare data obtained by means of a survey questionnaire with information on the same subject from administrative records. The latter are assumed to represent the standard against which the survey responses are to be judged, although it should be recognized that administrative records are themselves subject to error.[1]

The main objective of a record check when used in questionnaire design and evaluation is methodological--to determine whether the desired information can be obtained by a survey. Can respondents recall the events, can they report them with reasonable accuracy, and are they willing to do so? Subsidiary objectives include ascertaining which kinds of topics are better reported and which are not, and determining the appropriate reporting period for asking respondents to recall events.

There are two basic approaches to the conduct of a record check. Usually, a sample of persons with the desired characteristics or experiences is drawn from administrative records and an attempt is made to interview these individuals with a questionnaire designed to elicit responses that can be compared with information from the records. This approach is generally referred to as a reverse record check. The alternative is to select a sample of survey

[1]For a more complete discussion of matching techniques, see U.S. Department of Commerce (1980).

Principal Contributor: Richard W. Dodge

questionnaires and attempt to match them with administrative records so that answers to similar topics can be compared. This method has been called a forward record check.

In the development of questionnaires, the reverse record check has important advantages. It provides at reasonable cost a sample of persons possessing the characteristics or exhibiting the behavior that one may wish to study in a full-scale survey. This is especially desirable when the variable of interest occurs so rarely that screening the general population for eligible cases would be prohibitively expensive. In situations where the subject matter is unfamiliar to those responsible for the survey, or there is not much previous experience to draw on, a sample of persons known to possess the desired attributes can provide clues as to the proper way to phrase questions, or even to test whether the desired information can be usefully collected by a sample survey. The following discussion is focused on the reverse record check as an aid to questionnaire design. However, one of the examples at the end of the chapter is of a forward record check.

II. METHOD

A. Personnel and Skill Requirements

A record check is a labor-intensive procedure, both in transcribing the data from the records and in determining whether information available from both sources is a match. Depending on the type of records and arrangements with the recordholders, the data transcription may be done either by clerical personnel from the survey organization or by the recordkeeping agency.

The matching operation is done by the survey organization, and may require professional staff assistance in resolving problems with the matching, in addition to clerical personnel (if the matching is done by hand) or computer-experienced personnel (if the matching is done by computer). The choice between these two methods of matching may be based on the number of records--it probably would not be cost-effective to do a computer match with relatively few records.

In addition, the field work stage requires interviewers and all associated tasks of interviewer recruitment, interviewer training, etc. After the completion of the interviewing and matching phases, qualified professionals are needed to analyze the results.

B. Selection of Respondents

Respondents for the field test are selected in the record transcription phase of the project. The number of records selected is determined by available funds and the degree of precision required of the results. If the records are ordered chronologically, and this is an important element of the test, a systematic selection should be made; otherwise, a convenience sample may be sufficient. Since there will be a time lag between the occurrence of the event and the interviews, allowance should be made in sample selection for movers who have left the area, as well as for bad addresses and persons who can never be found.

C. Preparation

Before adopting the record check as part of the questionnaire design process, one should be aware of the issues that are likely to arise in implementing such a procedure. These issues need to be evaluated against the goals of the particular survey in order to determine whether a record check should be incorporated into the survey development plan.

Timing is of crucial importance in implementing a record check. The first step is to locate a record system containing the desired information and one from which a sample can be drawn. One system with the requisite number of cases would be preferable to drawing samples from several systems, but this may not be possible if a series of sequential record checks is planned or a variety of record systems is needed to test all the important variables. Proximity to the survey organization may be an important consideration in selection of a record system as a way of keeping costs down.

Before deciding on a particular source for record-check cases, there are other matters that need to be addressed. Obtaining permission to use administrative records may be time consuming, even if the records are open to the public. At the least, a letter describing the survey and the kinds of information needed from the records must be sent to the appropriate official under the signature of the head of the survey organization or other responsible person. Before obtaining permission, it would probably be advisable to determine how the records are organized--chronologically, by subject matter, geographically, etc.; whether they need to be reordered before a sample can be selected; what form the records take--paper copies of originals, computer printouts, microfiche, magnetic tape, etc; whether the sampling can be done by the survey organization or must be done by the record holders; whether any of the information on the records is confidential and therefore must be blanked before the sample is chosen. Depending upon the responses to these questions, the decision can be made whether to request formal permission to use the records. If the sample selection cannot be done by the survey organization, it is important to obtain an estimate of the cost of the work to be performed by the record holders, as well as an idea of the time they will need to select the sample and prepare the cases for follow-up in the field. The time element may be important in deciding whether to use a particular record system, especially if the record-keepers select the sample on a time-available basis, rather than on a predetermined schedule. No matter how promptly the sampling is done, there will inevitably be a time lag between when the events occur and when the field test takes place.

In addition to the basic information needed to locate respondents--name, address, telephone number--other descriptors should be identified to assist in matching cases obtained in the field with those in the original sample. The absence of adequate matching criteria, beyond the key items of interest, would be sufficient reason not to utilize a particular record system. Since respondents may report to interviewers similar or related events that were not part of the administrative record that caused the case to fall into sample in the first place, specific information about the event, in addition to its date of occurrence, may be as important as demographic characteristics of the respondent.

D. Operation

The field work stage of the record check should occur as soon as possible after the sample is selected. Interviewers should already have been recruited and trained. Depending upon how long it takes for events to be incorporated into the record system, a minimum of several months and possibly much more time will have elapsed since the target incident took place. It is therefore important to minimize further delays which would complicate efforts to find respondents and increase problems of recall.

Ideally, the purpose of a record check (which is to find out whether respondents will report a particular event) should not be revealed to either interviewers or respondents so as not to bias the results. However, it may be difficult and/or costly to maintain this stance in practice. Unless the questionnaire covers a great many other subjects, interviewers may notice that most respondents will report their involvement in a particular kind of event--attendance at plays or concerts, visits to physicians, victims of crime, etc. This could result in biasing the test because interviewer expectations could affect the results obtained. One way to minimize this possible effect is to supplement the sample with dummy cases, i.e., nearby addresses which would have a much lower probability of exhibiting the type of behavior being measured. However, this may greatly increase the cost of the test and might not entirely achieve its purpose. By not giving the interviewer the sample respondent's name, interviews would have to be administered to all potentially eligible persons in the household. Cases might also have to be sent back to the field for an explanation if there is no interview with the sample person, although this would nullify the attempt to disguise the survey purpose.

An explanation of the purpose of the survey should be prepared for interviewers to give as part of their introduction. A general statement which does not reveal the source of sample will probably satisfy most respondents. But some will press for more information--a telephone number to authenticate the survey auspices or an explanation of how they happened to be selected. In the latter instance, the survey designer must decide how far to go in revealing the source of the sample, although a candid response is usually the best policy.

Once the data have been collected, the critical process of matching respondent reports of particular events (doctor visits, crime incidents, etc.) with record information takes place. Many cases will be obvious matches, but there will be a substantial number of borderline situations where subjective judgment will enter in. For these cases, the matching criteria need to be clearly specified, as well as the degree of acceptable variation. However, it is difficult to specify guidelines for this activity because the number of variables used and the definition of what constitutes a matched case will vary according to the subject matter and the objectives of the study. The entire process should be completely recorded so that others can review the decisions made at this key stage.

A drawback of the reverse record-check method is that important aspects of the topic may not be covered by administrative records. For example, in studying the accuracy of reporting crime victimization by sampling police

records, it is obvious that only crimes that find their way into police record systems are included. Thus, one can ascertain which of the crimes sampled from police records a respondent failed to report, but one cannot draw conclusions about incidents reported to interviewers that were not in the administrative sample. The survey designer should be aware of this limitation of the reverse record-check method in questionnaire development.

E. Time Considerations

The record-check technique is a time-consuming one for several reasons. The initial research necessary to locate record systems and decide on the feasibility of their use, the process of securing permission for their use, and the record selection process all require the cooperation of outside parties. At each of these steps, unforeseen difficulties may be encountered and the time required to perform these tasks is unpredictable.

The data collection phase of a record check involves most of the same tasks as a formal or informal test of the same size. Because of the need to interview the person involved in the record event, however, tracking may be required to locate sample persons who have moved or for whom the original address information was inadequate. This process is time consuming and may add considerably to the time required to complete the data collection.

The final time-related aspect of a record check involves the record-matching process. The time requirements for this task depend on whether the matching is done by computer or clerically. (The latter is generally more time-consuming although in smaller record checks, of 100 cases or so, it may be faster to do the matching process clerically than to prepare computer specifications and write programs.) Timing will also depend on how precisely the matching criteria can be specified, which may affect the rate with which mismatches occur and the amount of time required for the professional staff to resolve these problems.

F. Cost Considerations

The cost factors for a record check are basically similar to a formal or informal test of the same size. However, the amount of time required to complete the data collection and matching phases affects the costs associated with the technique. The additional time required (by interviewers, clerical personnel, etc.) to complete these tasks is reflected in increments in the level of these costs.

There also may be costs associated with securing permission from the recordholders to use their materials or to have the records selected by personnel at the recordkeeping agency if the records cannot be released for confidentiality or other reasons. The geographic location of the recordholders may also be a cost factor if personal visits are required to examine the records and see how they are arranged, to select the sample, etc. The travel costs involved in accomplishing these tasks may be considerable.

G. Mode of Data Collection

Record checks are most frequently used in conjunction with personal visit surveys because the home address is normally a part of the administrative record. If telephone numbers are readily available, this interviewing mode can also be employed, although a small proportion of interviews may have to be done in person. Responses to a mail survey can be checked against administrative records, as in the examples of forward record checks described in Example 2, below, but the rate of return should be sufficiently high to guarantee the validity of the results. Using the mails in a reverse record check is considerably more risky because of the problems involved in tracking persons who have moved from the address on the administrative record.

III. EXAMPLES

A. Example 1: Crime Survey Tests

1. Introduction

The National Crime Survey (NCS) used reverse record checks for a number of purposes in preparing for the initiation of a nationwide survey in 1972. The procedure used was to draw a sample from police records of persons who had been victimized by certain crimes and then attempt to interview them with a questionnaire designed to elicit reports of victimizations.

A series of three reverse record-check tests were undertaken in preparation for the National Crime Survey. All were conducted by the Bureau of the Census under the sponsorship of the Law Enforcement Assistance Administration. The first test was held in Washington, D.C., with the sample of 484 cases drawn from Metropolitan Police Department records (U.S. Department of Commerce, 1970). Baltimore, Md., was the site of the second test which used a sample of 527 from Police Department files (Yost and Dodge, 1970). The final, and most elaborate, record check consisted of 620 cases of known victims selected from police records in San Jose, Calif. (Turner, 1972).

2. Objectives of NCS Record Checks

The most important objective of the NCS record checks was to aid in developing a victimization questionnaire by measuring the ability (or willingness) of crime victims to report to interviewers incidents of crime which had originally been reported to police authorities and recorded by them. A series of questions was formulated containing the elements of the kinds of crimes covered by the proposed survey--rape, robbery, assault, burglary, larceny, and motor vehicle theft. Persons selected from police files had been recent victims of one of these crimes. An underlying assumption was that questions that were successful in eliciting reports of incidents sampled from police records would also be appropriate for obtaining information about incidents not reported to the police--an assumption which could not be independently verified.

The questionnaire designed to achieve this objective was a combination screener and incident form, although varying versions of the questionnaire were used in each jurisdiction depending on the specific objectives of the

test and, in the case of Baltimore and San Jose, reflecting experience gained in earlier tests. The screener contained a series of questions, phrased in nontechnical language, intended to jog a respondent's memory about the kinds of crime which would eventually be included in the National Crime Survey. The incident form collected detailed information about each reported incident so that a match could be made with the sample cases from police files. In all all three NCS tests, crimes other than those selected for the record-check sample were reported by the respondents when they were interviewed. The information gathered on the questionnaire was generally sufficient to distinguish these additional incidents from the ones in the record check.

In addition to matching as many incident reports as possible, another objective was to ascertain the degree of correspondence between the survey's classifications of the crimes and those assigned by the police. An important related objective was to determine the ability of the respondent to report certain other facts about the incident that could be verified by the police record. These included such items as estimates of property loss, characteristics of the offender(s), and month of occurrence of the incident.

Other objectives were crucial to the development of the NCS. These included the length of the reference period to use in asking about crime incidents befalling respondents, the degree to which respondents moved ("telescoped") incidents into the reference period that occurred outside it (usually earlier), and the degree to which events, although located properly within the reference period, were not placed in the correct month.

3. Technical and Operational Considerations

a. Selection of the test sample from police records

The test samples were drawn soon after the close of the reference period about which respondents were asked to report their victim experience. This was not only because of anticipated memory decay, but, more importantly, because of the difficulty in locating victims of crime, especially violent crime, who appear to be a highly transient group. The success rate in finding and interviewing crime victims averaged about 66 percent for the three NCS record checks.

Direct access to police files in order to draw a sample was not possible in all three jurisdictions, so detailed sampling specifications had to be prepared for police personnel. To do this properly, it was necessary to know how the files of offense reports were organized, whether the files were computerized, what information was available about the incident, whether the initial police report contained more information than was in the computerized file and, if so, whether the police report could be made available. Where it was necessary for the police to draw the sample, the time schedule for the test had to allow for the police department's ability to fit this work in with their regularly assigned duties.

b. Information needed from police records

Sufficient information about incidents and victims had to be obtained from police records to facilitate a match between cases selected and cases

interviewed. Achieving this goal was complicated by police confidentiality requirements in one case and by the sparse amount of information on the computerized file in another case. For example, in Washington, D.C., the initial police reports were public documents and copies were readily available. However, the police had to select the sample because confidential material about incidents was filed with the police report. In Baltimore, copies of the police report were not available and identification information about victims and details of incidents had to be hand-copied from police reports after the sample was selected from computerized files. Knowledge of the victims' places of work, hours of work, and office telephone numbers obtained from police files proved extremely useful to interviewers in tracking down some difficult-to-reach respondents.

c. Field operations

Interviewer training stressed techniques for locating respondents, in addition to a thorough review of the content of the test questionnaire.

Although, as noted earlier, it is desirable to commence field activities as soon as possible after sample selection, one should avoid starting when only part of the sample has been chosen. The latter situation caused problems in the Washington test. Because of delays in the police selection of cases, interviewers were assigned cases on a flow basis. Since the police files were organized by month of occurrence of the incident, cases were assigned whenever a particular month's sample was selected. This proved to be inefficient because cases received in the latter part of an assignment were often for addresses in neighborhoods that had been visited earlier.

Although it was recognized that informing interviewers of the source of the sample cases and providing them with the names and addresses of victims could bias the results, there did not seem to be any reasonable alternative. Having the name of the victim made it possible to follow up many of the cases which could not be found at the initial address. Without the victim's name (and information relating to jobs held when that was available), completion rates would have been far lower in the Washington and Baltimore tests.

The San Jose record-check test was held under different circumstances in that it was conducted at the same time as a victimization survey of the general population. The general population sample was about 8 times larger than that in the record check. Thus, it was easier to mask the fact from interviewers that part of the workload came from police files. For both kinds of cases, interviewers were supplied with addresses, but not names. However, it was apparent to some interviewers that the record-check cases had distinctive identification numbers and that these households produced many more crime events than did the other households. Also, record-check cases were subjected to an office edit to ensure that the victims had been interviewed. If no filled questionnaire was found, the interviewers were then given names and other pertinent information and instructed to try to locate and interview the victims.

At first, it was thought undesirable for interviewers to tell respondents initially how their names had been selected for fear of biasing the results. However, the need to telephone many persons in advance to arrange an interview

usually required a more lengthy explanation of the purpose of the survey than was needed in a personal interview. Interviewers were instructed to inform respondents that their names had been selected from police records when asked directly or whenever the interviewers felt it was necessary to gain cooperation. This knowledge had no discernible impact on the substance of the interview or on the respondents' willingness to participate in the survey.

4. Results of Record-Check Tests

The principal finding of the three record-check tests for the National Crime Survey was that the crimes covered by the survey could be elicited to an acceptable degree by the questionnaire as it had evolved by the time of the San Jose test. The results from that test are shown in Table 1. With the exception of assault, the recall rate for the other major crimes was collectively above 80 percent. Evidence from each of the tests demonstrated that assault was the least well recalled (or reported) of the crimes. It was also apparent that aggravated assault, the more serious form of the crime, was better reported than simple assault. In addition, the closer the relationship of the victim to the offender, the less likely was an assault incident to be reported to an interviewer. Thus, assaults by strangers were well reported, but assaults by relatives were often not mentioned.

One important caveat in using crime incidents drawn from police records should be noted. Crimes reported to the police and subsequently reported in survey interviews undoubtedly differ from those that are never brought to police attention. In general, the former tend to be more significant and therefore more salient in respondents' minds. Questions which elicit reports of such events may provide an overestimate of what the level of recall would be for all crimes of a particular type.

Table 1. San Jose Reverse Record Check: Incidents Reported, by Type of Crime

Type of crime	Total cases interviewed	Incidents reported in survey	
		Number	Percent
Total	394	292	74.1
Rape	45	30	66.7
Robbery	80	61	76.3
Assault	81	39	48.1
Burglary	104	94	90.3
Larceny	84	68	81.0

Source: Turner, 1972: 6.

As a result of the record-check tests, several modifications were made in the final questionnaire. Initially, it was intended that the screening questions would indicate the specific crime involved and that interviewers would fill an incident form tailored to that crime. It soon became clear that the sole function of the screening questions should be to gather all the incidents that respondents were willing and/or able to report, but that no attempt should be made to classify crimes at that stage. To facilitate

the recall of incidents, the number of screening questions was increased and additional examples of incidents were incorporated into the question wording. Thus, the determination of which type of crime was involved (including those incidents which were not crimes or were out-of-scope for the survey) was made from the data collected on the incident report. For the regular survey, a single incident report was designed that could be used to record all incidents. Ultimately, the classification of incidents was done by computer.

Expansion of the incident report in the test phase was due, in part, to the need for extra information in order to be able to match back to police records and to be better able to classify the crime. In addition, more subjects of analytical interest were included on the final version of the questionnaire, such as characteristics of offenders, data on the nature of property taken and/or damaged, extent of personal injury sustained.

By the conclusion of the test phase, there was substantial agreement between the classification of incidents by the police and that stemming from the survey. Most remaining differences, in fact, seemed to be traceable to local crime definitions which varied from those employed by the FBI's Uniform Crime Reports, the standard used in the survey.

The record check demonstrated that a respondent's ability to recall whether an incident occurred was not appreciably better when a recall period of 6 months was used compared with one of 12 months. However, respondents were less accurate in placing an event in its proper month of occurrence when the recall period was 12 months. Since accurate placement of incidents in time was an important consideration in the survey, the 6-month period was chosen. A 3-month recall period would have resulted in greater accuracy, but would have required twice the sample size to achieve the same degree of reliability as the 6-month period.

The Washington record check documented the tendency of respondents to report events, which actually took place earlier, as having occurred within the recall period. A bounding interview was thus introduced in the main survey to control this tendency by establishing a time frame which can be used in the subsequent interview to edit out incidents occurring before the beginning of the recall period. Data from the bounding interview are not used in preparing NCS estimates. However, households that move into sample addresses in the second through the seventh times that the unit is in the sample are not bounded for their first interview. Reporting incidents that took place later as having occurred during the recall period is less common and can be minimized by conducting interviews as soon as possible after the end of the recall period.

B. Example 2: 1980 Census Tests

1. Introduction

Record checks were used in the development of the questionnaire for the 1980 Census of Population and Housing to test questionnaire content and census procedures. In the examples described here, responses to census test questionnaires were checked against data supplied by utility companies and units of government. These are examples of forward record checks, i.e., the sample

was selected from test questionnaires and responses were subsequently matched with independent records. The purpose of these record checks was not so much to develop question wording as to see whether reasonably accurate data could be collected on these subjects.

Some of the issues that were discussed in the National Crime Survey example were not relevant in these record checks. The selection of the geographic areas for the tests was based on criteria that were not related to the availability of records for checking test responses.

However, for the purpose of checking census reports of utility costs and mortgage status, the likelihood of a selected area having a good record system was extremely high. In addition, unlike crime victimization, utility costs and mortgage status are not rare events. These tests also did not encounter problems of confidentiality. Matching problems were not as great as long as the census address was the same as the billing address. Nonmatches on address were deleted, as were households that did not have a minimum number of months of service provided by the utility companies.

2. Reporting of Utility Costs

The purpose of including average monthly utility costs (electricity and gas, in this example) in the census test questionnaires was not only to obtain information on the accuracy of respondent reporting of these items, but also to study the impact of errors in the calculation of gross rent (contract rent plus utility costs) and shelter costs for homeowners.[2] Gross rent and shelter costs, rather than utility costs, are the items shown in 1980 census publications.

The record checks were held in Travis County (Austin), Texas (Fronczek, 1977) and Oakland, Calif. (Koons, 1979). Systematic samples of owners and renters for each type of utility were selected from filled questionnaires received in the mail-out/mail-back procedure. The Travis County test, which was restricted to the city of Austin, resulted in the following numbers of census cases matched to utility company records: 626 owner-electricity, 459 renter-electricity, 608 owner-gas, and 365 renter-gas. The comparable figures from the Oakland test were 667, 568, 652, and 475.

The results indicated that census responses were higher than the reports of the utility companies. The overreporting error in electricity cost was about 45 percent in Travis County and 48 percent in Oakland; for gas, the error for Travis County was 78 percent and for Oakland it was 33 percent.

---

[2]Shelter cost is a new concept in census data and includes the average monthly cost of mortgage payments (if applicable), utility payments, real estate taxes, and fire and hazard insurance. The 1980 census was the first in which utility cost information was collected from home owners.

Table 2 shows gross rent and shelter costs based on census reports and utility company bills.[3] This shows that the percent difference was lower for Oakland than for Travis. In part, this is due to the relatively low utility costs in Oakland with respect to the overall components of gross rent and shelter cost for owners. Although the errors in utility cost estimates were fairly large, the combined shelter cost estimates were thought to be accurate enough for most purposes.

Table 2. Gross Rent and Shelter Costs Based on Census Estimates and Utility Company Bills: Travis and Oakland

			Shelter costs for owners			
	Gross rent		Mortaged		Not mortgaged	
Characteristics	Travis	Oakland	Travis	Oakland	Travis	Oakland
a. Census report	$181.05	$196.53	$264.23	$342.18	$103.12	$121.14
b. Utility company report	165.76	189.16	236.85	333.35	81.96	113.43
c. Difference (a-b)	15.29	7.37	27.38	8.83	21.16	7.71
d. Percent net difference $\left(\frac{a-b}{b} \times 100\right)$	9.2	3.9	11.6	2.6	25.8	6.8

Source: Koons, 1979: 5.

3. Reporting of Mortgage Status

Travis County was also the site for a test of a question asking whether homeowners had any mortgage debt (Benedik, 1977).[4] This question was asked for the first time in the 1980 census. To test the ability of respondents to answer this item accurately, a record check was undertaken. A probability sample of 745 mortgaged and 570 nonmortgaged cases was studied. Certain assumptions had to be made to carry out this record check. An investigation of the mortgage documents was ruled out as too time consuming. It was then decided that an adequate check could be done by comparing the mortgaged and nonmortgaged properties as indicated on the census questionnaires with municipal tax records which indicated who was billed for the city tax, since tax bills were mailed directly to the lender in those cases where a loan was in force. Thus, the absence of a loan company number meant that the homeowner was billed for the tax and, therefore, the property was not mortgaged. A supplemental record check was carried out in those cases with inconsistent responses when the Census answer was compared with tax records, i.e., those

[3]The differences in the figures between the census reports and the utility company reports were due solely to differences in the costs of utilities; the other components of gross rent and shelter costs came from Census sources in both cities.

[4]The specific wording of the question was "Do you have a mortgage, deed of trust, contract to purchase, or similar debt on this property?"

without loan company numbers and a "mortgaged" response on the census and those cases with loan company numbers and a "not mortgaged" response on the census.

Table 3 summarizes the result of the entire record check. For the 745 mortgaged census reports, 705, that is, 95 percent, were mortgaged according to the records. For the 570 census questionnaires where the respondent reported no mortgage, 531, or 93 percent, were nonmortgaged.

The data showed a relatively high consistency in answering the mortgage status item correctly, sufficient to justify its placement on the census questionnaire. The response errors for mortgage status of 5 percent for mortgaged properties and 7 percent for nonmortgaged properties were judged to be within acceptable limits.

Table 3. Census Responses to Mortgage Status and Corresponding Record-Check Determinations

Record-check results	Number	Percent
Census questionnaire said property was mortgaged		
Total	745	100
Mortgage or similar debt	705	95
Mortgage	623	84
Contract for sale	82	11
Not mortgaged	40	5
Census questionnaire said property was not mortgaged		
Total	570	100
Mortgage or similar debt	39	7
Mortgage	14	3
Contract for sale	25	4
Not mortgaged	531	93

Source: Benedik, 1977: 9.

The principal difficulty with the mortgage question involved the 107 cases classified by the census as contracts for sale (contracts to purchase). These are equivalent to mortgages but do not have formal mortgage documents and are not required to be recorded. The additional record check of inconsistent responses revealed that for the 82 cases said to be mortgaged in the census, someone else paid the taxes and was the owner of record. On the other hand, the 25 contract-for-sale cases who claimed that their property was not mortgaged apparently did not consider that this arrangement came within the meaning of the census question. Among the 40 cases that reported mortgages, but for which no record of debt could be found in the record check, there is the possibility that a mortgage holder was a private individual or a lender not on the city tax list of lenders. Resources were not available to make additional investigation of these cases. The final census questionnaire (long form) contained the same question wording, but provided a separate answer category for contracts to purchase.

Chapter 11
Response Analysis Surveys

I. INTRODUCTION

This chapter will deal with a questionnaire development or evaluation technique often used with establishment surveys that are conducted by mail. The technique is known as a response analysis survey (RAS),[1] in which a sample of respondents in the mail survey are personally interviewed. A structured questionnaire from which the answers can be tabulated is used in the interview.

Questionnaire designers can use information, provided directly by respondents, to evaluate the reliability and validity of data that are currently being collected in a repetitive survey or that are proposed for collection in the future. When used prior to developing a new questionnaire or questions, the RAS seeks to determine what steps should be taken to obtain quality responses. For example, since establishment surveys frequently require extensive use of records by respondents, the RAS could determine how these records are kept. Then, questions could be designed to take advantage of the information provided in the records, thus making the job easier for respondents. When used in a repetitive establishment survey, RAS information can be used to refine the questionnaire. The fact that a questionnaire or reporting form may have been used for many years does not preclude the need for reviews and reevaluations. For example, the questions may need to be revised if the recordkeeping systems of the establishments change. Other changes that might prompt revisions include new definitions of items, the need to obtain new information, and new laws that affect the availability of some data.

In repetitive mail surveys, many questionnaire problems will surface informally as a result of refusals to respond by newly selected establishments, inquiries from respondents regarding definitions, requests for data, and so forth. However, it is assumed that other problems might be detected if interviewers and observers collected the data on a regular basis. An RAS allows both subjective and objective evaluation and analysis of continuing mail survey questionnaire items.

---

[1]This technique and accompanying terminology has been used by the Bureau of Labor Statistics for a number of years. It is also used by Statistics Canada and may be known by other names elsewhere.

Principal Contributor: Carol M. Utter

II. METHOD

A. Personnel and Skill Requirements

If the survey is repetitive, the staff working on it are probably best able to decide what questions should be asked in the RAS. If the survey is new, experienced survey and questionnaire designers and other subject matter experts should be used. The actual interviews should be conducted by personnel who have had experience in interviewing or who have, at least, had training in interviewing techniques. In addition to interviewing ability, it is important that the interviewers know enough about the survey under review or the purpose of the new survey so they can answer any question that may arise during the interview.

B. Selection of Respondents

The size of the subsample of panel members selected for interview is usually determined by the resources available. It also depends on the sample size of the survey being evaluated. The degree to which the results of the RAS will be disaggregated for purposes of analysis will also influence the size of the subsample. For example, if the results are to be analyzed by two or three subgroups (e.g., industry, size of business, region, form type, etc.) sufficient numbers of establishments within these groups must be selected for the RAS. If the purpose of the interviews is to evaluate or review a repetitive survey, the sample selection process must ensure that all types of respondents are represented. A systematic sample of panel members will meet this criterion. If the RAS results will be used to make statistical inferences, the establishments need to be selected according to a stratified sampling design developed for this purpose.

C. Preparation

Designers of an RAS questionnaire to review an existing survey should first avail themselves of all the information about the survey that may have accumulated; e.g., proposals for new survey initiatives, questions regarding definitions, misinterpretations of instructions, requests for changes, and results from other types of formal or informal reviews of the survey. Obviously, this type of information will not be available to assist in designing an RAS questionnaire for use prior to developing a questionnaire for a new survey. However, examples of similar RAS forms can serve as guides when either evaluation or development is the purpose of the RAS. Each data item collected or proposed for collection in the main survey should be covered by a set of questions regarding that item. For example, are records maintained on the specific item; how long does it take for the respondent to assemble the information; must part of it be estimated sometimes, always; are the instructions clear?

There is a limit, however, to how many questions can be asked effectively in the RAS. Experience has shown that the attention span of respondents is less than 1 hour, so the validity of the answers from the later part of the interview may be compromised if the interview is longer. If too many questions are required because of the number of questions on the main survey, the questionnaire could be designed to cover only part of the data items or to cover

some of them in less detail. Another possibility is to split the RAS sample in two or more parts and ask about different questions in each part of the sample.

In most respects, drafting of the RAS questionnaire follows the same path taken for other questionnaires. There is one caveat, however. It is most important that respondents in a continuing survey are not "turned off" by the interview. Sensitive questions should be considered carefully since one does not wish to lose the respondent's cooperation in the main survey. Additionally, when structuring questions to review a repetitive survey, any indication of which answer is the "correct" one based on the main survey's definitions should be avoided, so as not to imply that the respondent is currently reporting incorrectly. Thus, the RAS, in addition to providing valuable information about the survey, can also serve as a public relations vehicle, particularly if the only other contact with the panel of establishments is through the mail. Interviewees are often pleased to be a part of the review process.

Once a draft of the questionnaire has been prepared, three or four of the most knowledgeable subject matter persons should try it out on a small number of respondents in local establishments. A debriefing session following these initial interviews should be helpful in identifying and correcting any problems with the flow and interpretation of the questions. The tests will also provide a check on the time required to complete the interview. (See Chapter 9 for further discussion of debriefing sessions.)

D. Operation

Before the interviews can be conducted, the person (or persons, in large business establishments) who will be the respondent must be identified and contacted. If the survey is new, a letter addressed to the firms selected for the RAS should precede a telephone call to set up an appointment. The letter can merely introduce the purpose of the survey and the visit. In the telephone call (within 1 or, at most, 2 weeks after the letter was mailed) an interview should be requested with a person who is knowledgeable in the subject to be covered by the RAS. In small firms, this person will most likely be the owner or manager; in large firms, the interviewer would ask to speak to the personnel director, payroll clerk, chief engineer, etc., depending on the subject of the proposed interview. In a continuing survey, the mail survey form probably identifies a contact person, together with a telephone number. However, this person may not be the optimal respondent. For example, a clerk may routinely fill out a monthly mail survey form and be listed as the contact, but the department head may be a better choice for the interview. Sometimes more than one person may be designated to answer questions since various parts of the RAS may require different kinds of expertise.

After a time has been agreed on for the interview, data collection proceeds as in any other interviewer-administered survey.

The responses obtained in the RAS are tabulated and carefully evaluated to determine the following types of information. Were all questions answered? Are the answers consistent--that is, do answers to one question contradict

the answers to other similar questions? What suspected inaccuracies in the data collected in the repetitive survey can be identified? A report analyzing the results of the RAS would be expected to lead to recommendations for or against the main survey and, probably, further testing of main survey questionnaires or forms. In this respect, the RAS is no different than most research projects.

E. Time Considerations

The time required to complete the RAS is, again, largely dependent on the availability of personnel to do the planning, execution, and analysis of survey results. Six months may be considered a minimum time frame under the best of circumstances.

F. Cost Considerations

The largest cost factors of the RAS are the salaries of the interviewers and the travel costs. The costs of the forms design, reproduction of questionnaires and other materials, training of interviewers, and data processing services must also be considered. If the RAS is for a repetitive survey, substantial cost savings can be realized if regular staff members conduct the interviews in addition to their other duties. Many of the other costs may not be separable from the main survey under review.

G. Mode of Data Collection

This technique is most appropriate for developing a mail survey questionnaire or revising an existing mail survey questionnaire. By using a different mode of data collection for the RAS than is used or proposed for the main survey (i.e., personal visit interviews versus mail questionnaires), additional information can be obtained that would otherwise not be available.

III. EXAMPLE: RAS OF RESPONDENTS IN THE CURRENT EMPLOYMENT STATISTICS SURVEY

The Bureau of Labor Statistics cooperates with 51 state employment security agencies in collecting data each month on employment, hours, and earnings from a sample of about 200,000 establishments in all nonagricultural activities including government. From these data, a large number of monthly economic series are compiled for the United States, for each of the 50 States and the District of Columbia, and for most of the metropolitan areas. The data include series on total employment, production or nonsupervisory worker employment, number of women workers, average hourly earnings, average weekly hours, and average weekly overtime hours (in manufacturing) in considerable industry detail. The survey, known as the Current Employment Statistics (CES) Survey, has been conducted by the Bureau of Labor Statistics since 1915, but has undergone many changes through the years. It is conducted entirely by mail and cooperation by employers is voluntary.[2]

---

[2]For a more complete discussion of the CES Survey, see Chapter 2 of the BLS Handbook of Methods (1982).

The questionnaire used for this survey since 1930 (BLS-790 series) is a "shuttle" schedule; that is, the schedule is submitted monthly for 12 months by the respondent, the information is copied from it by BLS and then it is returned to the respondent for use again the following month. There are several variants of the schedule with detailed instructions and definitions designed to meet the specific problems of different industries. Most of the schedules ask for the entry of five data items from a summary of the establishment's payroll so that only about 10 minutes of the respondent's time are required each month. (See Figure 1, a facsimile of the BLS-790 C used for manufacturing establishments, at the end of this chapter.)

Following the report of the National Commission on Employment and Unemployment Statistics (1979) which recommended a number of changes in several major U.S. statistical systems, the BLS embarked on a program to modernize the CES Survey. As a first step in this effort, the Bureau conducted an RAS with representatives of establishments that are regular reporters in the program. Since this was the first formal attempt in 25 years to review the record-keeping practices of the employers who cooperate in the CES survey, a lot of unanswered or partially answered questions had accumulated.

A small task force of the most experienced staff members was assigned to develop a draft questionnaire for the RAS. The task force asked for and received suggestions from staff working on the survey in different states. They assembled these into a first draft and circulated it to all interested parties.

Comments and further refinements followed. These were incorporated in a second draft which also benefited from the services of a forms design expert. The draft was then informally tested through interviews with nine respondents. As a result of problems identified in this small test, several questions were eliminated because the draft questionnaire was too long. Other questions were tightened and some multiple choice checkoff items were added. The final questionnaire required an average of 30 minutes of a respondent's time to complete. (Examples of sections of the questionnaire are provided in Figure 2.) It included questions on how employers maintain their records (manually or computerized); whether outside contractors are used to summarize records; the types of occupations for which summary statistics are available; whether separate pay records, hours of work records, and/or personnel records are kept for occupational groups; how much time elapses before specific payroll summaries are available; whether it is necessary to estimate parts of the data reported; whether establishments have provisions for paid sick leave, paid holidays, paid vacations, premium overtime, and shift differential and whether these items can be separately identified on their payroll records. Several questions were directed at the accuracy of the responses and respondents' perceptions of Form BLS-790--e.g., did they have difficulty understanding and following the instructions, and what changes, if any, would they like to see made in the collection forms.

A subsample of current panel members (those who had responded at least once during the 6 months prior to sample selection) was chosen for interviews in four states--Florida, Massachusetts, Texas, and Utah. The interviews were conducted by staff of the CES cooperating state agencies in those states. In addition, BLS regional office and national office staff interviewed

representatives of large companies which maintain special reporting arrangements with the Bureau. Response to the RAS was extremely high; only 3.5 percent of respondents contacted refused to be interviewed.[3] Altogether, 1,071 interviews were completed--an adequate sample to represent the views of 180,000 employers who provide the monthly reports. Interviewers were asked to prepare short unstructured post-interview reports containing impressions of the interview and any pertinent additional information. These reports indicated that many panel members of the CES program actually welcomed the opportunity to be heard. (See Chapter 9 for further discussion of the use of interviewer evaluations.)

The RAS took about 1 year to complete and cost $200,000 (1981 dollars). Each phase of the survey took about 3 months--the development of the questionnaire, OMB clearance plus planning and arranging for the field work and interviewer training, data collection, and the tabulation and preparation of the preliminary analysis. A final report on the RAS was published a year later.

The RAS results revealed that about two-thirds of employers cooperating in the CES Survey maintain computerized payrolls, but only about one-third maintain payroll summaries according to the current definitions used by BLS. (Since CES Survey participation is voluntary, this may not reflect the state of payroll summaries of all employers.) Nevertheless, three-fourths of the respondents claimed that they spend less than 20 minutes each month in preparing Form BLS-790. However, deviations from stated definitions and concepts were not uncommon in the reported data. For example, only about one-fourth of employers said that they maintained records on the number of women workers on their payrolls, but 94 percent regularly report a number for women workers (as requested) to BLS each month. On the other hand, only 8 percent of employers said that they "estimated" this number. The rest counted first names on the payroll list or relied on "personal knowledge" to derive the number reported. (The "personal knowledge" answer was used by several employers of hundreds of employees.) Another example relates to manufacturing employers who did not use the stated definitions in classifying and reporting production workers. Contrary to instructions, 16 percent included supervisors and 14 percent included administrative and clerical personnel in their production worker counts, and over half included janitorial services not related to production processes. Many employers, who were aware of the instructions, noted that these employees were not separately identified in their payroll records, so it was not possible to exclude them. Newly established types of production workers, e.g., computer technicians engaged in manufacturing a product, are evidently considered to be nonproduction workers by three-fourths of employers who employ such workers in production activities.

Armed with the results of the RAS, the staff proposed several substantive changes to the Forms BLS-790. These included changes in format and definitions as well as the collection of new data items and the elimination of old

---

[3]Establishments for which the monthly forms were prepared in a central office location outside of the boundaries of the four States were not contacted by State personnel. Many of these were then contacted by BLS regional office and national office personnel.

ones. The proposed new data items include total payrolls and hours for all workers and for part-time workers in service and trade establishments; however, elimination of the reporting of payrolls and hours for nonsupervisory workers was suggested. A new "short form" to be used by part of the panel was also proposed to help reduce reporting burden. The revised Forms BLS-790, at this writing, are being tested through another small sample of respondents to ascertain that the changes do, in fact, produce timely and accurate reporting of employer payroll information.

References

Andrews, Frank M., and Withey, Stephen B. 1976. Social Indicators of Well-Being: Americans' Perceptions of Life Quality. New York: Plenum Press.

Atkinson, J. 1968. A Handbook for Interviewers. Government Social Survey (No. M136). London: HMSO.

Atkinson, Tom. 1977. "Is Satisfaction a Good Measure of the Perceived Quality of Life?" Paper presented at the Annual Meeting of the American Statistical Association.

Babbie, Earl R. 1973. Survey Research Methods. Belmont, Calif.: Wadsworth Publishing Company.

Bayton, James A. May 1978. Qualitative Analysis of a Proposed New Form for Application for a Social Security Number. (Unpublished report prepared for the Social Security Administration.)

Belson, William A. 1981. The Design and Understanding of Survey Questions. London: Gower Publishing Co., Ltd.

Benedik, R. S. June 30, 1977. "Results of the Mortgage Status Record Check," 1976 Census of Travis County, Texas, Results Memorandum No. 20, U.S. Bureau of the Census.

Bercini, Deborah, and Massey, James. 1979. "Obtaining the Household Roster in a Telephone Survey: The Impact of Names and Placement on Response Rates." American Statistical Association, Proceedings of the Social Statistics Section, pp. 136-140.

Biderman, Albert D.; Cantor, David; and Reiss, Albert J., Jr. 1982. "A Quasi-Experimental Analysis of Personal Victimization Reporting by Household Respondents in the National Crime Survey." Paper presented at the Annual Meeting of the American Statistical Association.

BLS Handbook of Methods, Vol. I, Bulletin 2134-1, December 1982.

Bohannan, Paul. 1964. Africa and Africans. Garden City, N.Y.: Natural History Press.

Bradburn, Norman M.; Sudman, Seymour; and Associates. 1979. Improving Interview Method and Questionnaire Design. San Francisco: Jossey-Bass Publishers.

Cahoon, Lawrence; Kniceley, R. Maurice; and Shapiro, Gary M. 1980. "Informational Needs for Current Demographic Survey Design With Discussion of Key Redesign Research Projects." Paper presented at the Annual Meeting of the American Statistical Association.

Cannell, Charles F., and Robison, Sally. 1971. "Analysis of Individual Questions." Working Papers on Survey Research in Poverty Areas. Ann Arbor: Survey Research Center, U. of Michigan.

Cannell, Charles F.; Lawson, S. A.; and Hausser, D. L. 1975. A Technique for Evaluating Interviewer Performance. Ann Arbor: Survey Research Center, U. of Michigan.

Cannell, Charles F.; Marquis, Kent H.; and Laurent, Andre. 1977. "A Summary of Studies of Interviewing Methodology, 1959-1970." Vital and Health Statistics. Series 2, No. 69. DHEW Pub. (HRA) 77-1343.

Converse, Jean, and Schuman, Howard. 1974. Conversations at Random: Survey Research as Interviewers See It. New York: John Wiley and Sons, Inc.

Dillman, Don A. 1978. Mail and Telephone Surveys: The Total Design Method. New York: John Wiley and Sons.

Duffy, Thomas M. 1981. "Organizing and Utilizing Document Design Options." Information Design Journal, vol. 2/3 and 4, pp. 256-265.

Fronczek, Peter J. March 18, 1977. "Accuracy of Reports of Average Monthly Utility Costs (Gas and Electricity) for Owner and Renter Households," 1976 Census of Travis County, Texas, Results Memorandum No. 14, U.S. Bureau of the Census.

Gibson, Christina O.; Schapiro, Gary M.; Murphy, Linda R.; and Stanko, Gary J. 1978. "Interaction of Survey Questions as It Relates to Interviewer Respondent Bias." American Statistical Association, Proceedings of the Section on Survey Research Methods, pp. 251-256.

Givens, Jimmie D., and Moss, Abigail J. 1981. "Redesigning the National Health Interview Survey's Data Collection Instrument." Paper presented at the Annual Meeting of the American Public Health Association.

Glynn, Thomas. July 7, 1978. "Report on Interviewer and Respondent Replies to Questions about the Consumer Expenditure Survey," U.S. Bureau of the Census (unpublished memorandum).

Goode, William J., and Hatt, Paul K. 1952. Methods in Social Research. New York: McGraw-Hill Book Co.

Hainer, Peter. Dec. 1979. "Census Definitions and the Politics of Census Information: Working Misunderstandings in a Northeast Urban Poor Black Setting." Paper presented at the Annual Meeting of the American Anthropological Association.

Henson, R.; Cannell, Charles; and Lawson, S. 1973. Effects of Interviewer Style and Question Form on Reporting of Automobile Accidents. Ann Arbor: Survey Research Center, U. of Michigan.

Hoinville, G.; Jowell, R.; and Associates. 1978. Survey Research Practice. London: Heinemann Educational Books Ltd.

Human Services Research Inc. October 22, 1980. A Research Report of the Development of Methods To Collect Willingness-To-Pay Data for the 1980 Hunting and Fishing Survey.

Hyman, Herbert; Cobb, J.; Feldman, J.; and Stember, Charles. 1954. Interviewing in Social Research. Chicago: U. of Chicago Press.

"Income Survey Development Program: 1979 Research Panel Documentation." Dec. 1982. Information on the availability of this document can be obtained from Mr. Stuart Weisman, National Technical Information Service, 5285 Port Royal Road, Springfield, Va., 22161 (703-487-4807).

Jabine, Thomas J., and Rothwell, Naomi D. 1970. "Split-Panel Tests of Census and Survey Questionnaires." American Statistical Association, Proceedings of the Social Statistics Section, pp. 4-13.

Kahn, Robert, and Charles Cannell. 1957. The Dynamics of Interviewing. New York: Wiley.

Koons, David A. July 20, 1979. "Accuracy of Reports of Utility Costs for Occupied Households," 1977 Census of Oakland, California, Results Memorandum No. 37, U.S. Bureau of the Census.

Laurent, Andre; Cannell, Charles F.; and Marquis, Kent H. 1972. "Reporting Health Events in Household Interviews: Effects of an Extensive Questionnaire and a Diary Procedure." Vital and Health Statistics, Series 2, No. 49, DHEW Pub. 72-1049.

Liebow, Elliot. 1967. Tally's Corner. Boston: Little, Brown.

Marquis, Kent H. 1971. "Purpose and Procedures of the Tape Recording Analysis." Working Papers on Survey Research in Poverty Areas. Ann Arbor: Survey Research Center, U. of Michigan.

Morton-Williams, Jean. 1979. "The Use of 'Verbal Interaction Coding' for Evaluating a Questionnaire." Quality and Quantity, vol. 13.

Moser, Claus, and Kalton, Graham. 1972. Survey Methods in Social Investigation. New York: Basic Books.

National Commission of Employment and Unemployment Statistics. 1979. Counting the Labor Force. Washington, D.C.: Government Printing Office.

Office of Management and Budget, Office of Information and Regulatory Affairs. June 1983. "Contracting for Surveys," Statistical Policy Working Paper 9.

Olson, Janice, and Vaughan, Denton R. 1982. "The 1979 Income Survey Development Program Research Panel Test of Attitude Measures: Some Preliminary Results." ISDP Staff Paper, Office of Research and Statistics, Social Security Administration.

Payne, Stanley. 1951. The Art of Asking Questions. Princeton, NJ: Princeton University Press.

Pelto, Pertti J. 1970. Anthropological Research: The Structure of Inquiry. New York: Harper Row.

Reiss, A. J., Jr. 1961. Occupations and Social Science. New York: Free Press of Glencoe.

Richardson, S.; Dohrenwend, Barbara; and Klein, D. 1965. Interviewing: Its Forms and Functions. New York: Basic Books.

Rothwell, Naomi D. Feb. 11, 1974. "Results of Debriefing Interviewers on the Consumer Expenditure Survey," U.S. Bureau of the Census (unpublished memorandum).

_____. Feb. 19, 1974. "Some Information About Consumer Expenditure Survey Respondents Obtained During Reinterview and Interviewer Evaluation of Respondents," U.S. Bureau of the Census (unpublished memorandum).

_____. 1980. "Discussion of 'The Effect of the Question on Survey Responses.'" American Statistical Association, Proceedings of the Section on Survey Research Methods, pp. 40-42.

Scherr, Marvin G. 1980. "The Use of Focus Group Interviews To Improve the Design of an Administrative Form: A Case Study at the Social Security Administration." American Statistical Association, Proceedings of the Section on Survey Research Methods, pp. 347-352.

Schuman, Howard, and Presser, Stanley. 1981. Questions and Answers in Attitude Surveys: Experiments on Question Form, Wording, and Context. New York: Academic Press.

Sirken, Monroe D. Nov. 1972. Designing Forms for Demographic Surveys. Laboratory for Population Statistics Manual Series No. 3.

Statistics Canada. Dec. 1980. Social Concepts Directory for Statistical Surveys. Ottawa: Department of Supply and Services.

Social Science Research Council. 1975. Basic Background Items for U.S. Household Surveys. New York: Social Science Research Council.

Stack, Carol. 1974. All Our Kin: Strategies for Survival in the Black Community. New York: Harper & Row.

Sudman, Seymour and Bradburn, Norman M. 1974. Response Effects in Surveys: A Review and Synthesis. Chicago: Aldine.

_____. 1982. Asking Questions: A Practical Guide to Questionnaire Design. San Francisco: Jossey-Bass Publishers.

Turner, Anthony G. 1972. The San Jose Methods Test of Known Crime Victims. Washington, D.C.: National Criminal Justice Information and Statistics Service, Law Enforcement Assistance Administration, U.S. Department of Justice.

Turner, Charles F., and Martin, Elizabeth, Eds. 1984. Surveying Subjective Phenomena. (2 vols.) New York: Russell Sage Foundation and Basic Books.

U.S. Department of Commerce, Bureau of the Census. June 10, 1970. "Victim Recall Pretest, Washington, D.C.: Household Survey of Victims of Crime" (unpublished memorandum).

U.S. Department of Commerce, Office of Federal Statistical Policy and Standards. June 1980. "Report on Exact and Statistical Matching Techniques," Statistical Policy Working Paper 5.

Vaughan, Denton R., and Lancaster, Clarise G. 1979. "Income Levels and Their Impact on Two Subjective Measures of Well-Being: Some Early Speculations From Work in Progress." American Statistical Association, Proceedings of the Social Statistics Section, pp.169-174.

_____. 1980. "Applying a Cardinal Measurement Model to Normative Assessments of Income: Synopsis of a Preliminary Look." American Statistical Association, Proceedings of the Section on Survey Research Methods, pp. 47-52.

Wright, Patricia W. 1981. "Informed Design of Forms." Information Design Journal, vol. 2/3 and 4, pp. 151-178.

Ycas, Martynas, and Lininger, Charles A. Nov. 1981. "The Income Survey Development Program: Design Features and Initial Findings." Social Security Bulletin, vol. 44, pp. 13-19.

Yost, Linda R.,and Dodge, Richard W. Nov. 30, 1970. "Household Survey of Victims of Crime, Second Pretest, Baltimore, Md." U.S. Bureau of the Census (unpublished memorandum).

☆ U.S. GOVERNMENT PRINTING OFFICE: 1985-461-105/20225

Reports Available in the Statistical Policy Working Paper Series

1. Report on Statistics for Allocation of Funds
   GPO Stock Number 003-005-00178-6, price $2.40

2. Report on Statistical Disclosure and Disclosure-Avoidance Techniques
   GPO Stock Number 003-005-00177-8, price $2.50

3. An Error Profile: Employment as Measured by the Current Population Survey; GPO Stock Number 003-005-00182-4, price $2.75

4. Glossary of Nonsampling Error Terms: An Illustration of a Semantic Problem in Statistics (A limited number of copies are available from OMB.)

5. Report on Exact and Statistical Matching Techniques
   GPO Stock Number 003-005-00186-7, price $3.50

6. Report on Statistical Uses of Administrative Records
   GPO Stock Number 003-005-00185-9, price $5.00

7. An Interagency Review of Time-Series Revision Policies (A limited number of copies are available from OMB.)

8. Statistical Interagency Agreements (A limited number of copies are available from OMB.)

9. Contracting for Surveys (Available through NTIS Document Sales, PB-83-233-148.)

10. Approaches to Developing Questionnaires (Available through NTIS Document Sales, PB-84-105055.)

Copies of these working papers, as indicated, may be ordered from the Superintendent of Documents, U.S. Government Printing Office, Washington, D.C. 20402 (202-783-3238) or from NTIS Document Sales, 5285 Port Royal Road, Springfield, VA 22161 (703-487-4650).